NO
PAPER
RIGHTS

REIMPRESSION
INTERDITE

DESIGNING AND CONDUCTING HEALTH SYSTEMS RESEARCH PROJECTS

Corlien M. Varkevisser
Indra Pathmanathan
Ann Brownlee

Health Systems Research Training Series

Volume 2 Part 2

IDRC

The International Development Research Centre is a public corporation created by the Parliament of Canada in 1970 to support research designed to adapt science and technology to the needs of developing countries. The Centre's activity is concentrated in six sectors: agriculture, food, and nutrition sciences; health sciences; information sciences; social sciences; earth and engineering sciences; and communications. IDRC is financed solely by the Parliament of Canada; its policies, however, are set by an international Board of Governors. The Centre's headquarters are located Ottawa, Canada. Regional offices are located in Africa, Asia, Latin America, and the Middle East.

The World Health Organization is a specialized agency of the United Nations with primary responsibility for international health matters and public health. Through this organization, which was created in 1948, the health professions of some 165 countries exchange their knowledge and experience with the aim of making possible the attainment by all citizens of the world by the year 2000 of a level of health that will permit them to lead a socially and economically productive life.

By means of direct technical cooperation with its Member States, and by stimulating such cooperation among them, WHO promotes the development of comprehensive health services, the prevention and control of diseases, the improvement of environmental conditions, the development of health manpower, the coordination and development of biomedical and health services research, and the planning and implementation of health programs.

Other volumes in the HSR Training Series

Volume 1: Promoting health systems research as a management tool (IDRC-286e)
Ann Brownlee

Volume 3: Strategies for involving universities and research institutes in health systems research (IDRC-288e)
Ann Brownlee, Lilia Duran Gonzales, and Indra Pathmanathan

Volume 4: Managing health systems research (IDRC-289e)
Indra Pathmanathan

Volume 5: Training of trainers for health systems research (IDRC-290e)
Indra Pathmanathan and N.I. Nik-Safiah

Il existe également une version française de cette publication.
La edición española de esta publicación también se encuentra disponible.

HSR Training Series

Volume 2: Designing and Conducting Health Systems Research Projects

Part II: Data Analysis and Report Writing

The Technical Working Group

Ann Brownlee (United States)
Lilia Duran Gonzales (Mexico)
German Gonzales (Colombia)
Yvo Nuyens (Belgium)
Indra Pathmanathan (Malaysia)
Annette Stark (Canada)
Patrick Twumasi (Ghana)
Corlien M. Varkevisser (The Netherlands)

IDRC-287e.2

Part II: Data Analysis and Report Writing

Designing and Conducting Health Systems Research Projects

Health Systems Research Training Series
Volume 2

Corlien M. Varkevisser
Indra Pathmanathan
Ann Brownlee

Jointly published by the Health Sciences Division of the International Development Research Centre, Ottawa, Canada, the Programme on Health Systems Research and Development of the World Health Organization, Geneva, Switzerland. This series was translated into Spanish by the Health Services Development Program of the Pan American Health Organization.

© International Development Research Centre 1991
PO Box 8500, Ottawa, Ontario, Canada K1G 3H9

Varkevisser, C.M.
Pathmanathan, I.
Brownlee, A.

IDRC, Ottawa, Ont. CA
WHO, Genève, CH

IDRC-287e.2

Designing and conducting health systems research projects. Pt. 2, Data analysis and report writing. Ottawa, Ont., IDRC 1991. xxi + 167 p.: ill. (Health systems research training series ; v. 2)

/Health surveys/, /research projects/, /project design/, /health services/ - /training courses/, /data processing/, /data analysis/, /risk/, /case studies/, /tests/, /participatory research/, /research results/, /research reports/, /writing and editing/, /references/.

UDC: 613.001.5

ISBN: 0-88936-585-7

A microfiche edition is available.

The views expressed in this publication are those of the author and do not necessarily represent those of the International Development Research Centre or the World Health Organization. Mention of a proprietary name does not constitute endorsement of the product and is given only for information.

Abstract

This is the second volume of a five-volume Health Systems Research (HSR) Training Series which has been compiled by a Technical Working Group supported by IDRC and WHO. Each volume is directed toward a particular target group and each addresses specific aspects of the HSR process. Volume 2, in modular format, is the pivotal one which deals step-by-step with the development of an HSR proposal and field testing (Part I), and with data analysis and report writing (Part II). Course participants will select, preferably in advance of the course, priority health problems particular to their own situations that cannot be solved unless more information is collected. In most cases, a team of course participants will then carry out the planned research alongside their regular duties. A second workshop is then scheduled to provide information on data analysis, report writing, and utilization of results. This volume will be of interest to all target groups and especially to those health-care managers and researchers who wish to conduct HSR projects.

The other volumes in the series are: volume 1, which focuses on the need to promote the use of HSR as management tool and reviews strategies for promoting HSR among policymakers and senior managers; volume 3, a review of strategies that can assist universities or research institutes to initiate and implement multidisciplinary HSR programs; volume 4, a course outline in modular format designed to provide research managers with the skills for managing a program of HSR; volume 5, a course outline in modular format, designed to assist those whose primary responsibility is organizing and conducting training courses for the relevant target groups.

The series is designed to support a program of essential national health research. Users are encouraged to critically examine the materials and to choose or adapt them to their particular needs.

Résumé

Ce volume est le deuxième d'une collection de cinq volumes de formation à la recherche sur les systèmes de santé (RSS) qui ont été rassemblés par un groupe de travail technique financé par le Centre de recherches pour le développement international et l'Organisation mondiale de la santé. Chaque volume est destiné à un groupe particulier et chacun porte sur certains aspects de la recherche sur les systèmes de santé. Le volume 2, sous forme modulaire, est le volume central qui expose, étape par étape, la manière de formuler une proposition de RSS et de la mettre à l'essai (partie I) et d'analyser les données et de rédiger un rapport (partie II). Les participants des cours choisiront, de préférence avant les cours, les problèmes de santé qui sont prioritaires pour eux dans leur travail et qui ne pourront être résolus sans un supplément d'information. Dans la majorité des cas, les participants sont groupés en équipes pour faire la recherche planifiée en plus de leurs tâches ordinaires. Un second cours est ensuite organisé sur l'analyse de données, la rédaction de rapports et l'utilisation des résultats. Ce volume intéressera tous les groupes cibles et surtout les gestionnaires de soins de santé et les chercheurs de ce domaine qui veulent exécuter des travaux de recherche sur les systèmes de santé.

Les autres volumes de la collection sont les suivants: le volume 1 traite de la nécessité de promouvoir la RSS comme outil de gestion. Y sont décrites les stratégies propres à cette promotion auprès des décideurs et des cadres supérieurs. Le volume 3 vise à aider les chercheurs de formation universitaire qui travaillent dans des universités ou des instituts de recherche et qui veulent promouvoir des programmes multidisciplinaires de RSS et y participer. Le volume 4 est un guide de gestion d'un programme de RSS. Le volume 5 aidera les personnes chargées d'organiser et de donner des cours de formation aux divers groupes cibles.

Ces cinq volumes ont pour but d'appuyer la création d'un programme national de recherche essentielle en santé. Les personnes qui s'en serviront sont incitées à les examiner d'un oeil critique et à en tirer ce qui répond à leurs besoins ou y répondrait après adaptation.

Resumen

Este es el segundo de cinco volúmenes de una serie de capacitación sobre Investigación de Sistemas de Salud (ISS), compilada por un Grupo de Trabajo Técnico que recibió el apoyo del Centro Internacional de Investigaciones para el Desarrollo (CIID) y la Organización Mundial de la Salud (OMS). Cada volumen está dirigido hacia un grupo particular y trata de aspectos específicos del proceso de ISS. El Volumen 2, en formato modular, es un elemento fundamental que trata progresivamente del desarrollo de una propuesta de ISS y su prueba sobre el terreno (Parte I). Asimismo, se trata en este volumen el análisis de datos y la redacción de informes (Parte II). Los participantes del curso seleccionarán, preferentemente con antelación al curso, problemas de salud prioritarios, específicos de sus propias situaciones e imposibles de resolver hasta que no se recopile más información. En la mayoría de los casos, un equipo de participantes del curso llevará a cabo la investigación planificada conjuntamente con sus deberes regulares. A continuación se programa otro seminario para proporcionar información sobre análisis de datos, redacción de informes y utilización de resultados. Este volumen será de interés para todos los grupos específicos y especialmente para los administradores de establecimientos de salud e investigadores que deseen realizar proyectos de investigación sobre sistemas de salud.

Los otros volúmenes en la serie son: volumen 1, centra su atención en la necesidad de promover el uso de ISS como instrumento de gestión. Asimismo, describe las estrategias para promover la ISS entre ejecutivos y gerentes principales; volumen 3, concebido para ayudar a los investigadores con educación universitaria que trabajan en universidades o institutos investigativos que deseen promover y participar en programas multidisciplinarios de ISS; volumen 4, guía para la gestión de un programa de ISS; volumen 5 servirá de ayuda a aquellos cuya responsabilidad primaria sea organizar y dictar cursos de capacitación para los grupos meta pertinentes.

La serie está diseñada para apoyar un programa esencial de investigación sobre salud a nivel nacional. Se exhorta a los usuarios a examinar críticamente los materiales y/o adaptarlos a sus necesidades particulares.

ACKNOWLEDGMENTS

The present volume, *Designing and Conducting Health Systems Research Projects*, has its roots in the course materials developed in the early 1980s by the Project for Strengthening Health Delivery Systems (SHDS), at the request of the WHO Regional Office for Africa (AFRO).

These materials were very popular and have been widely applied in workshops in Western and Central Africa, as well as in other parts of the world, to train health staff in developing and implementing problem-oriented research proposals. Nevertheless, modifications were necessary. It was felt that the content should be more adapted to meeting information needs of decision-making at the different levels of the health system and that a larger variety of research methods than offered in the original course should be presented. The present modules 1-17 are heavily adapted or new versions of the original SHDS modules. Furthermore, the need was recognized to support course participants beyond the point of just developing a research proposal, through the phases of fieldwork, data analysis, and report writing. Therefore, an additional set of modules was developed (Modules 18-32).

This adaptation and extension took place in the WHO Subregional Office III, Harare, and the Public Health Institute of Kuala Lumpur, Malaysia, first independently and soon in close collaboration through WHO's Health Systems Research (HSR) group at headquarters.

In Harare, the Joint HSR Project (a joint enterprise of WHO and the Royal Tropical Institute, Amsterdam, supported by the Netherlands Ministry for Development Cooperation), which promotes HSR in all countries of the southern African region, developed its set of modules with 10 researchers from the region. The Malaysian core group that developed the course material included two Malaysian scientists and a Sri Lankan statistician. These modules were used in numerous workshops in southern African countries, Malaysia, and in other countries and regions from 1988 to 1991, and were revised several times.

Since early 1989, when IDRC and WHO HQ took the initiative to support the development of this two-part volume, the different sets of modules have been gradually merged and further developed. Corlien M. Varkevisser (Zimbabwe), Indra Pathmanathan (Kuala Lumpur), and Ann Brownlee (formerly SHDS), who also did the final editing, are responsible for the present version. However, numerous others have contributed as well. First of all, the other members of the Technical Committee, who provided valuable advise and communicated their experiences with training in HSR in other parts of the world; then all those who participated in the development of the modules: L. Omondi (Botswana); M.E. Sebatane and T.K. Makatjane (Lesotho); P. Chimimba and L. Msukwa (Malawi); Maimunah Abdul Hamid and K. Mariappan (Malaysia); A. Kitua and E. Savy (Seychelles); C. Sivagnanasundram (Sri Lanka); G. Tembo (Zambia); R. Munochiveyi, P. Taylor, and G. Woelk (Zimbabwe); R. Peeters (Belgium); and, finally, M.W. Borgdorff and L. Bijlmakers (WHO Subregional Office III, Harare), who assisted substantially in the writing or rewriting of the different versions. Richard Hayes, Betty Kirkwood, and Tom Marshall were so kind to allow publication of materials used in the master of science course in community health in developing countries at the London School of Hygiene and Tropical Medicine for Modules 28 and 29.

A major input into this process of module development has been provided by the participants and facilitators, named and unnamed, in the various courses that have been part of the exercise. Finally, a great deal of effort was devoted to word processing, editing, formating, proofreading, and all the other painstaking tasks involved in publishing by another group of dedicated people.

To all of the above, thank you.

92-05437/

CONTENTS

Foreword xi

General Introduction xv

Introduction to This Volume xvii

Module 21: Orientation to the Workshop on Data Analysis and Report Writing 1

Module 22: Description of Variables, Part I 15

Module 23: Cross-Tabulation of Quantitative Data 33

Module 24: Analysis of Qualitative Data 47

Module 25: Description of Variables, Part II 63

Module 26: Choosing a Significance Test 77

Module 27: Determining Differences Between Groups: Part I — Analysis of Unpaired Observations 91

Module 28: Determining Differences Between Groups: Part II — Analysis of Paired Observations 109

Module 29: Measuring Associations Between Variables: Regression and Correlation 119

Module 30: Measures of Risk in Case-Control Studies 131

Module 31: Report Writing 143

Module 32: Promoting the Utilization of Findings 159

FOREWORD

The ultimate goal of any national health-development process is to enable its people to reach a level of health that at least enables them to participate actively in the social and economic life of the community in which they live. To attain this objective, existing health systems must be redirected to achieve equitable reallocation of resources for health - total coverage, increased accessibility to primary health-care services, and effective referral to secondary and tertiary levels of care. It is also relevant to develop appropriate mechanisms to promote effective community participation in the promotion and maintenance of health.

Such redirection of health systems may require changes in health-care planning and government policy; in the organization and administration of health and related services; in the financing and budgeting of systems and procedures; and in the selection and application of appropriate technology.

To effect the necessary changes, countries must decide on the best approaches to adopt. This requires detailed and accurate information on needs, possibilities, and consequences of recommended actions. Such information is often lacking, inadequate, or unreliable. As a result, decisions are based on assumptions and unjustified conclusions and often result in inappropriate policy choices, the consequences of which are only discovered after implementation.

Research is a systematic search for information and new knowledge. It serves two essential and powerful purposes in accelerating advances in health. First, basic or traditional research is necessary to generate new knowledge and technologies to deal with major unresolved health problems. Second, applied research is necessary to the process of identifying priority problems and to designing and evaluating policies and programs that will be of the greatest health benefit, using existing knowledge and available resources, both financial and human.

These two purposes together, in what has now been defined as essential national health research, must catalyze the generation of new knowledge and the application of existing knowledge, an essential link to equity in development.

During the past decade, concepts and research approaches to support health development have evolved rapidly. Many of these have been described by specific terms such as operations research, health services research, health manpower research, policy and economic analysis, applied research, and decision-linked research. Each of these has made crucial contributions to the development of Health Systems Research (HSR) but their limited and highly focused approaches to problem solving have resulted in their being integrated within the scope of HSR while at the same time describing their unique contribution to health in development.

HSR is ultimately concerned with improving the health of a community, however defined, by enhancing the efficiency and effectiveness of the health system as an integral part of the overall process of socioeconomic development.

The aim of HSR is to provide health managers at all levels with the relevant information that they need to solve the problems they are facing. The participatory nature of such research is one of its major characteristics. It is argued that the involvement of all parties - the community, health-care managers and decision-makers, and researchers - in the definition of the problem helps to focus the investigation and to enrich the quality of the data collected. Similarly, participating in all stages of the research is essential if feasible and acceptable solutions to problems are to be implemented and sustained at community, district, regional, or national level.

Because HSR addresses health problems in the broad context of social, economic, and community development, research inputs from many different disciplines are required. These include demography, epidemiology, health economics, policy and management sciences, social and behavioral sciences, statistics, and some aspects of the clinical sciences. Each of these disciplines has developed specialized research approaches in its efforts to provide information that will support health development, but it is increasingly evident that the problems that are addressed by HSR require a combined input from many disciplines and especially that researchers from these specialized fields need to acquire the skills to work together in multidisciplinary teams.

The main characteristics of HSR are

- Its focus on priority problems in health;
- Its participatory nature;
- Its action orientation;
- Its integrated, multidisciplinary approach;
- Its multisectorial nature;
- Its emphasis on cost-effectiveness;
- Its focus on practical, timely solutions; and
- Its iterative nature that allows for evaluation of the impact of planned change and consequent revision of action plans and health policy.

Although its methodologies can be applied to similar problems in different countries, the findings and solutions to these problems are unlikely to be the same because of differences in cultural, social, economic, and political realities. This is one of the strong arguments in support of a national core of persons trained in HSR whose orientation and plan of work is guided by the country's agenda of essential national health research.

With progressive development, the uses of HSR are becoming more widely appreciated. As a result, it is being integrated into and applied in special areas of management such as quality assurance, technology assessment, and resource management.

Because the capacity for HSR is small, especially in developing countries, it is not surprising that, over the last few years, a series of training programs has been organized or funded by many agencies, including the International Development Research Centre (IDRC), the Pan American Health Organization (PAHO), the World Health Organization (WHO), and the US Agency for International Development (USAID).

As well, several international health programs have given high priority to capacity building for HSR.

- The UNICEF *Special Program on National Capacity Building for Child Survival and Development* aims "to strengthen awareness, knowledge, and skills for operations research using the health systems approach to promoting inquisitiveness and self-reliant approaches to identify pressing problems at the community level and find practical solutions for them."
- The overall goal of the *Network of Community Oriented Educational Institutions for Health*

Sciences is "to improve the relevance of health professions education by enhancing the ability of graduates to help identify and solve the problems of communities in which they serve ... using as framework a new system of partnerships among universities, governments and communities, the focus of which is a program of essential national health research."

- The *International Health Policy Program* is planning to develop health-policy research and training centres, whose role will be to facilitate and coordinate the "synthesis of policy-relevant research, dissemination of such research, capacity building in health policy analysis, and technical assistance for policy analysis and research."

- The *International Clinical Epidemiology Network* (INCLEN) supports the development of clinical epidemiology units (CEUs) in medical schools in developing countries. The role of CEUs is to provide leadership in the application of quantitative measurement principles (drawn from clinical epidemiology, biostatistics, health economics, and health social science) in the research, educaiton, and service responsibilities of the medical school.

- The Danish International Development Agency (DANIDA) has been supporting a series of inter-regional training workshops for research managers in HSR and, since 1987, the *Joint Project of the World Health Organization* and *The Netherlands Ministry for Development Cooperation - The Royal Tropical Institute* is involved in a process for capacity building for HSR in 14 countries of southern Africa.

All these and many more initiatives in capacity building for applied research received, in 1990, a strong political, moral, and intellectual backing in the recommendations of the Commission on Health Research for Development. In its *Agenda for Action*, the Commission recommends

> That building and sustaining research capacity be integrated as a key objective and powerful instrument for all health and development investments. Primary commitment must come from developing-country governments to accord priority and provide sustained financial support. Strong international re-enforcement is also needed. International exchange and interaction can do a great deal to help strengthen the capacity of developing-country researchers and institutions.

Within the broader context of the Commission's recommendations, three major challenges for the future development of HSR can be identified:

- How to enhance the demand for HSR;

- How to strengthen national capacities in HSR; and

- How to institutionalize the efforts into a sustainable process.

It is with these challenges in mind that this *Health Systems Research Training Series* was developed.

Annette Stark, Associate Director
Health Systems Research
Health Sciences Division
International Development Research Centre

Yvo Nuyens, Programme Manager
Health Systems Research and
Development
World Health Organization

GENERAL INTRODUCTION

A recent review of Health Systems Research (HSR) workshops sponsored by IDRC concluded that, although IDRC's objectives had been met, training materials should be revised and expanded to meet the needs of specific groups and to guide the development of follow-up sessions. In a related action, the WHO Global Advisory Group on HSR concluded that building and sustaining national capacities for HSR was a major issue to be addressed in program activities. It was specifically recommended that these activities must include components to "evaluate and revise training materials periodically and to support training programs at different levels of the health systems."

As a result of these recommendations, representatives of IDRC, PAHO, and WHO met in Ottawa in October 1988 to review past and current initiatives and to propose future activities. The group recognized that, if training in HSR is to have an impact on improving health and health care, it is necessary to clarify the context and stages of development of an effective HSR process within a given country. It was further decided that specific target groups for orientation and training in HSR should be selected and appropriate training strategies developed to strengthen the research capacity of countries, based on their specific needs and capabilities in HSR.

To achieve this goal, a technical working group was established and given the mandate to define and coordinate the development of a basic set of training materials for each of five identified target groups. The framework consisted of:

- A definition of the target group;
- A description of the entry competence or entry characteristics of the target group;
- The expected outcome behaviour, including skills and attitudes;
- The appropriate training strategies and training context; and
- The available training materials.

The deliberations and effort of the Technical Working Group have resulted in these five volumes of materials. Users are encouraged to become familiar, generally, with the entire set and then to selectively implement a program of training, research, planning, and health-care policy based on their country's needs.

Volume 1: Promoting Health Systems Research as a Management Tool

For Decision-makers
This document focuses on the need to promote the use of HSR as a management tool among decision-makers. Based on an analysis of experience in developing countries in the last decade, it presents an overview of how HSR can lead to better decisions and how the development of an effective research program can be fostered at country level. In addition, it provides descriptions of specific strategies for promoting HSR among policymakers and senior managers that have been used successfully in a number of settings.

Volume 2: Designing and Conducting Health Systems Research Projects

Part I - Proposal Development and Fieldwork
Part II - Data Analysis and Report Writing
Course participants, who may include concerned citizens, health workers, researchers, and health decision-makers from the provincial or even national level, will select priority health problems particular to their own situations that cannot be solved unless more information is collected. Preferably, the topics will have been selected before the training starts (see Volume 1), but they may need more specification. In most cases, a team of course participants will then carry out the planned research alongside their regular duties (Part I). A second workshop is then scheduled to provide information on data analysis, report writing, and utilization of results (Part II).

This volume is the pivotal one that deals specifically with the development of research proposals of a participatory nature (community/health-care manager/researcher) and, subsequently, with the implementation of the field study and the analysis and dissemination of study results. In this context, it is also of interest to junior researchers and those persons in universities and other training facilities who wish to operationalize HSR.

Volume 3: Strategies for Involving Universities and Research Institutes in Health Systems Research

For Senior Researchers and Academic Staff
This volume is designed to assist university-trained researchers located in universities or research institutes who wish to promote and participate in multidisciplinary programs of HSR. This volume will be of particular interest to those who wish to integrate the concepts of HSR into existing health and social science degree programs and to promote the development of student theses in the area.

Volume 4: Managing Health Systems Research

For Research Managers
The research managers for whom this volume is intended include managers of research institutes, academic departments, and agencies that have a function in processing research applications and in funding and coordinating research projects. The training should enable managers to facilitate their institutions' or organizations' contribution to and support of the development of HSR in the country as well as the utilization of research in improving the health of the people.

Volume 5: Training of Trainers for Health Systems Research

For Trainers and Facilitators
Experienced researchers are not necessarily experienced teachers. Moreover, few of them have experience in the organization and training of participants for whom research is a secondary responsibility and who have limited time to read or engage in research activities.

For training in HSR to be effective, experienced researchers need to acquire competence in the training approaches that have been successfully developed and used during the past few decades for training health personnel in a variety of important topics related to health.

Trainers and facilitators include those whose primary responsibility is organizing and conducting training courses for the different target groups and those who assist trainers in conducting courses.

INTRODUCTION TO THIS VOLUME

This publication is meant to be used in combination with Part I: Proposal Development and Fieldwork. Part I consists of 20 training modules which, step by step, support course participants in the development of a research proposal and provide useful guidelines for its implementation.[1]

The present Part II, Data Analysis and Report Writing, consists of 12 modules. These training modules on data analysis and report writing, to a still larger extent than those on the development of a research proposal, can be used in a flexible way, depending on:

- the educational level and research experience of the course participants;
- the type of study conducted and type(s) of data collection techniques used; and
- the state in which the data are at the onset of the data analysis and report writing workshop.

If participants have some previous training in research methodology and statistics, and/or some research experience, presentations of modules may be short. In this case the purpose of presenting is mainly to refresh the participants' memories and to guide them towards correct application of appropriate analysis procedures and tests. Some modules may be combined or shortened.

If participants have neither training nor experience in research, the presentation of the materials in the modules may have to be restricted to the bare essentials required to handle the data that has been collected. Under these circumstances presentations may take longer and should include ample opportunity for asking questions and for classroom exercises.

"Bare essentials" that could be considered include:

- Module 21 (Orientation to the workshop).
- All of Modules 22, 23 and 25 (Description of variables and cross tabulation).
- Module 24 (Analysis of qualitative data), especially the first two parts.
- Module 26 (Choosing a significance test). Concentrate on parts I, II and III, explaining the rationale for significance tests and how they work, but deal only briefly with part IV, the actual choosing of a significance test, if the groups are not likely to use more than the chi-square and/or the t-test.
- Module 27 (Determining differences between groups: Analysis of unpaired observations). Either the t-test or the chi-square test or both.

---

[1] The reader is referred to the introductory pages to Part I of Volume 2 for information on the other four volumes that the Health Systems Research Training Series, the rationale for publishing the series, acknowledgements, and a specific introduction on how to use both Part I and Part II of Volume II: Designing and Conducting Health Systems Research Projects.

- Module 30 (Measures of risk: incidence, risk, relative risk and odds ratio). Concentrate on unpaired observations.

- All of Modules 31 and 32 (Report writing and Promoting the utilization of findings).

Depending on the types of studies that participants have conducted and the analyses their data require, the scope of the presentations can be expanded (more on statistical tests, for example, or more on analysis of qualitative data), or the sequence changed (Module 24 may be presented before module 23 if participants have mainly qualitative data).

- Usually the first half of the workshop (one week) is devoted to the finalization of data processing and to data analysis. All modules related to analysis (21-30) are presented during this week.

  Timing of these presentations has to be done carefully. Modules 21-24 can be presented before data processing has been completed. Module 25 on statistical description (measures of dispersion, standard deviation, standard error) should wait until the teams have finished processing their data, as the module is a direct introduction to statistical tests.

  Only when the data have been properly processed, and when participants are well underway with the preparation of cross-tabulations, should the modules that present various statistical tests be presented.

- The second half of the workshop concentrates on report writing, drafting of recommendations, and presentation and discussion in plenary of the main findings and the recommendations arising from the studies. In this week there are usually only two presentations: on report writing (Module 31) and on the utilization of research findings (Module 32). The last module is best presented just before the participants draft the summary of findings and the recommendations of their studies.

An example of a schedule for a 2-week course on data analysis and report writing is presented on the following pages.

If the level of participants is high and if the data have been satisfactorily processed before reconvening for the data analysis and report writing workshop, it may be possible to shorten the period of the workshop to 1 or 1½ weeks.

EXAMPLE OF A COURSE SCHEDULE
(as used in southern Africa)

Designing and Conducting HSR Projects: Data Analysis and Report Writing

Date/Time	Session	Responsible Person(s)
Monday		
0800 - 0830	Opening remarks	Course Coordinator
0830 - 0915	Presentation and discussion of preliminary results	Group 1
0915 - 1000	Presentation and discussion of preliminary results	Group 2
1000 - 1030	Tea	
1030 - 1115	Presentation and discussion of preliminary results	Group 3
1115 - 1200	Presentation and discussion of preliminary results	Group 4
1200 - 1230	**Module 21: Orientation to the workshop on data analysis and report writing**	Facilitator
1230 - 1400	Lunch	
1400 - 1530	Project Group Work	
1530 - 1600	Tea	
1600 - 1700	**Module 22: Description of variables, part I**	Facilitator
1700 - 1800	Project Group Work	
Tuesday		
0800 - 0900	**Module 23: Cross-tabulation of quantitative data**	Facilitator
0900 - 1300	Project Group Work (including tea)	
1300 - 1400	Lunch	
1400 - 1500	**Module 24: Analysis of qualitative data**	Facilitator
1500 - 1800	Project Group Work (including tea)	
Wednesday		
0800 - 1300	Project Group Work (including tea)	
1300 - 1400	Lunch	
1400 - 1500	Optional: Presentations of main results of group work: revised objectives, main cross-tables, results of qualitative analysis	All 4 Groups
1500 - 1600	**Module 25: Description of variables, part II**	Facilitator
1600 - 1800	Project Group Work (including tea)	

Thursday		
0800 - 0900	**Module 26: Choosing a significance test** (20 min.), followed by **Module 27, part I and II** (t-test)	Facilitator
0900 - 1300	Project Group Work (including tea)	
1300 - 1400	Lunch	
1400 - 1500	**Module 27, part III** (Chi-square test)	Facilitator
1500 - 1800	Project Group Work (including tea)	
Friday		
0800 - 0900	**Module 28 or 29 or 30** (depending on the needs of the groups)	Facilitator
0900 - 1300	Project Group Work (including tea)	
1300 - 1400	Lunch	
1400 - 1500	**Module 28 or 29 or 30** (depending on the needs of the groups)	Facilitator
1500 - 1800	Project Group Work (including tea)	
Saturday		
0800 - 1300	Project Group Work (including tea)	
Monday		
0800 - 0900	**Module 31: Report writing**	Facilitator
Rest of day	Project Group Work	
Tuesday		
Whole day	Project Group Work	
Wednesday		
0800 - 0830	**Module 32: Promoting the utilization of findings**	Facilitator
Rest of day	Project Group Work	
Thursday		
0800 - 1300	Project Group Work	
1300 - 1400	Lunch	
1400 - 1445	Presentation of results, recommendations, and plan of action	Group 1
1445 - 1530	Presentation of results, recommendations, and plan of action	Group 2
1530 - 1600	Tea	
1600 - 1645	Presentation of results, recommendations, and plan of action	Group 3

1645 - 1730	Presentation of results, recommendations, and plan of action	Group 4
1730 - 1800	Closing Ceremony	

Friday

0800 - 1300	Project Group Work to finalize reports
1300 - 1400	Lunch
1400 - 1445	Evaluation of the HSR Training Course
1445 - 1730	Project Group Work to finalize reports

Saturday (if necessary)

0800 - 1300	Finishing touch to reports

Health Systems Research Training Series

Volume 2, Part II: Data Analysis and Report Writing

Module 21:

ORIENTATION TO THE WORKSHOP ON DATA ANALYSIS AND REPORT WRITING

Module 21: ORIENTATION TO THE WORKSHOP ON DATA ANALYSIS AND REPORT WRITING

OBJECTIVES OF THE WORKSHOP

At the end of this workshop you should be able to:

1. **Identify** and define the basic concepts and procedures required for data analysis and interpretation.
2. **Analyze** and interpret the data collected for the research project which you developed during the first workshop and draw conclusions related to the objectives of your study.
3. **Write** a clear and concise research report and a summary of the major findings and recommendations for each of the different parties interested in the results.
4. **Present** the major findings and the recommendations of your study to policy-makers and managers and to the subjects of your research. Work together with them to finalize the recommendations.
5. **Prepare** a plan of action for the dissemination of the findings and recommendations (if required) for additional research.

I. **Review of the Field Experience**

II. **Introduction to the Workshop**

III. **Tasks to be Completed During the Workshop**

Review and Finalization of Data Processing

Data Analysis

Report Writing

Presentation of Summary of Findings and Recommendations

Drafting a Plan for the Implementation of the Research Results

I. REVIEW OF FIELD EXPERIENCES

Implementing your planned project proposal must have been a big challenge to you. There must have been a number of unexpected obstacles as you became involved in your fieldwork, but some successes as well. If everything went according to plan, you have collected your data; you have processed a large part of it, if not all; you have completed some of your analysis; and you have written a preliminary report on the experiences and results of your fieldwork. Your practical experiences in conducting the project are invaluable. Sharing them with others in this workshop will be a very useful exercise, as both your mistakes and successes can provide valuable lessons for the future.

Before providing an overview of the focus and activities of this workshop, we would like to spend some time listening to the experiences of each of the research groups.

EXERCISE: Presentations of Field Experiences

Present the preliminary report that your group prepared at the end of your field experience, following the guidelines given in **Module 20**. Answer any questions other participants or facilitators may pose at the end of your presentation.

Each group will have approximately 20 minutes for its presentation and then 20-30 minutes for questions and discussion.

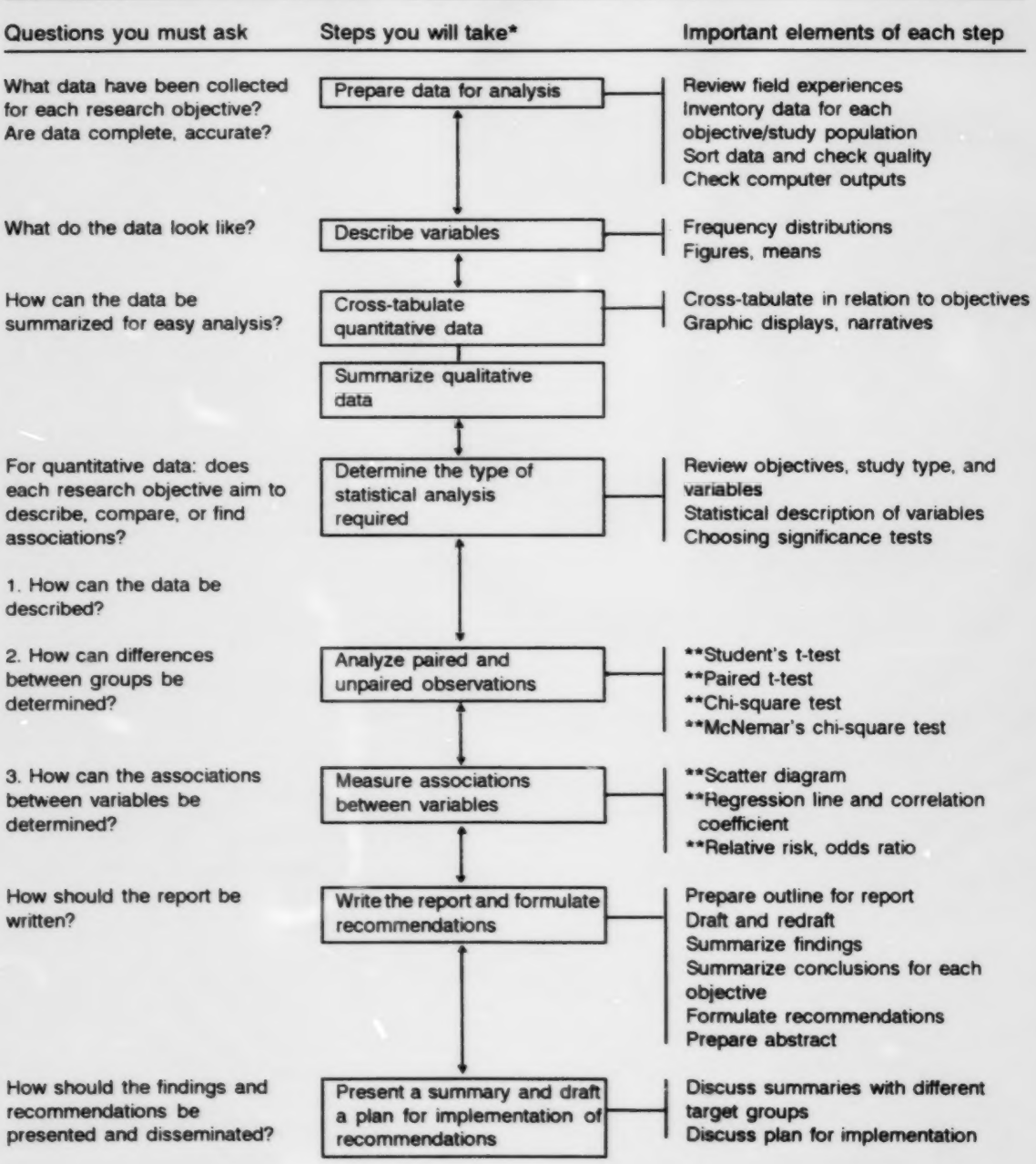
II. INTRODUCTION TO THIS WORKSHOP

This workshop is a follow-up of the workshop in which you developed your proposal. Now you have the major task ahead of fully analyzing the data you brought with you from the field and writing your research report. The report should contain feasible and useful recommendations based on the findings of your study, concerning how to solve the problem investigated.

As in the first workshop there will be presentations, group work sessions and a few plenaries, but in this workshop, group work will take up most of the time. The presentations will be concentrated in the first week, which will be devoted to data analysis. The second week will be fully reserved for report writing, with only two presentations to guide you. Toward the end of that week an important plenary is planned in which each group will present a summary of its main findings and recommendations. A selected group of policymakers and managers who requested the study or have a direct interest in the topic will be invited to comment on your presentation during that plenary.

The modules for this workshop cover several major tasks, which are schematically presented in the diagram on the next page of this module. This diagram is presented again at the beginning of each subsequent module, to indicate which task is the focus of the presentation. We will briefly discuss each of these tasks in this session.

Figure 21.1. Steps in data analysis and report writing.


Questions you must ask
Steps you will take*
Important elements of each step
What data have been collected for each research objective? Are data complete, accurate?
Prepare data for analysis
Review field experiences
Inventory data for each objective/study population
Sort data and check quality
Check computer outputs
What do the data look like?
Describe variables
Frequency distributions
Figures, means
How can the data be summarized for easy analysis?
Cross-tabulate quantitative data
Cross-tabulate in relation to objectives
Graphic displays, narratives
Summarize qualitative data
For quantitative data: does each research objective aim to describe, compare, or find associations?
Determine the type of statistical analysis required
Review objectives, study type, and variables
Statistical description of variables
Choosing significance tests
1. How can the data be described?
2. How can differences between groups be determined?
Analyze paired and unpaired observations
**Student's t-test
**Paired t-test
**Chi-square test
**McNemar's chi-square test
3. How can the associations between variables be determined?
Measure associations between variables
**Scatter diagram
**Regression line and correlation coefficient
**Relative risk, odds ratio
How should the report be written?
Write the report and formulate recommendations
Prepare outline for report
Draft and redraft
Summarize findings
Summarize conclusions for each objective
Formulate recommendations
Prepare abstract
How should the findings and recommendations be presented and disseminated?
Present a summary and draft a plan for implementation of recommendations
Discuss summaries with different target groups
Discuss plan for implementation

* These steps need not be in the sequence in this diagram. The sequence may be adjusted according to the needs of the workshop.
** These elements are optional and may be omitted if not relevant in a particular course.

III. TASKS TO BE COMPLETED DURING THE WORKSHOP

Review and Finalization of Data Processing

Although we trust that all groups have put great effort into the processing of their data, some adjustment and elaboration may be required. This is normal in research: each new step forward may be followed by half a step backward. Even in an advanced stage of data analysis you may still have to regroup and reprocess some of your data.

Before beginning data analysis, it is extremely important to check whether data processing has been carried out in such a way that information:

- is easy to handle; and
- has been checked for mistakes that may have crept in during data collection.

You, therefore, have to ask yourself the following questions:

- **Have the data been sorted appropriately?** Have questionnaires and checklists been numbered in the most convenient way? Can major categories of informants be clearly distinguished to facilitate comparison on relevant variables, as required by your research objectives?

- **Have quality checks been performed** on all data for completeness and consistency of information? Look at **Module 13** for measures to be taken in case of incompleteness or inconsistencies.

- **Have all data master sheets been filled in** with your quantitative data (or if you are using a computer, have all data been entered)? Do the total number of responses match the total number of respondents for each variable?

- **Have all qualitative data been categorized** as far as possible? If applicable, has coding been completed? (See **Module 13** for post-categorizing of open-ended questions.) Note that for qualitative data the collection, ordering, summarizing, and analysis are, in principle, intertwined. This will be taken up in detail in **Module 24**.

- If you used a computer to process your data, **check the frequency counts for each variable** in the questionnaire. Also check the computer cross-tabulations. Details on how to do this are given in **Annex 21.1**.

Before reviewing the data processing procedures, it is strongly advised that you make an INVENTORY of all data available for each OBJECTIVE. This is especially important if the data required have been collected using different collection tools.

Example:

Data sources for Objective 3: "Detection of weaknesses in the functioning of MCH services, explaining low utilization of delivery care."

- Questionnaire for mothers, Questions 12, 15-19, 23
- Focus group discussion with health staff, topics 3 and 4
- Observations included in checklists

Such an inventory will help you to organize data analysis better and, later, report writing.

Data Analysis

When beginning data analysis, we should consider which of our data are quantitative and which are qualitative.

Quantitative Data

Quantitative data are expressed in numbers and they are usually presented in frequency tables. From your data master sheets you can easily derive totals for each variable/question, count the number of different answers obtained, and present the information in a frequency table (**Module 22**).

When analyzing quantitative data it is important to reconsider the aim of our study. Is it to:

- **Describe variables?**

  For example: the distribution of teenage pregnancies in a certain population.

- **Look for differences between groups?**

  For example: differences between old settlers and newcomers in a certain area, with respect to income or health status.

- **Determine associations between variables?**

  For example: the association between work satisfaction of nurses and the number of staff meetings over the past year.

Cross-tabulations are the appropriate tool to summarize and analyze the data (**Module 23**).

In **Modules 25-30** some more advanced statistical concepts for the analysis of quantitative data are introduced. After frequency distributions and different types of cross-tabulations have been made, the type of statistical analysis required has to be determined. This includes further description of variables (**Module 25**) and selection of appropriate significance tests (**Module 26**).

The most common significance tests are:

- Student's t-test and the chi-square test to determine differences between groups if observations are unpaired (**Module 27**).

- The paired t-test and McNemar's chi-square test to determine differences between groups for paired observations (**Module 28**).

For measuring associations between variables the concepts of regression and correlation (**Module 29**) and odds ratio (**Module 30**) are introduced.

Throughout the process of data analysis, it is important to keep in mind that **our findings should provide answers to our research questions and thus meet our research objectives. We will eventually want to draw conclusions and make recommendations for action, based on these findings.**

Qualitative Data

You may remember that we obtain qualitative data through:

- **open-ended questions**, not precategorized, in questionnaires with predominantly quantifiable data;
- **loosely structured interviews** with open-ended questions directed at key informants (individuals or groups);
- **focus group discussions** on selected issues, with lists of points to guide the discussions; and
- **observations** to describe individual or group behaviour.

As you will remember from the "reasons for smoking exercise" in **Module 13**, the answers to open-ended questions may be:

- **listed**;
- **categorized** (based on your research objectives and common sense, combining the answers that belong together in some 4 to 6 categories, rarely more);
- **labelled or coded**;
- **inserted**, using these codes, **in your master sheets, or in the computer**; and
- **counted**, like other quantitative data.

However, if you are also interested in the CONTENT of each individual answer, for example because you want to start an anti-smoking campaign in which you will address different categories of smokers, you will further analyze the content of the answers for each category.

Note:

The major characteristic of analysis of qualitative data is that we analyze them **IN WORDS**, rather than in numbers.

Often it is useful to summarize qualitative data in diagrams, flow-charts, or matrices which help us in our analysis. **Module 24** will deal with the analysis of qualitative data in more detail.

Report Writing

You will be expected to go home with a completed report of your research. It will have the following components:

1. INTRODUCTION, covering a statement of the problem, some relevant data to explain the context, and literature review.
2. OBJECTIVES.
3. METHODOLOGY section with information on how you have collected your data (study type, variables, data collection tools); when and where (sample size, sampling procedures); how you have analyzed the data; and possible weaknesses in the collection and analysis.

4. FINDINGS AND CONCLUSIONS.

5. DISCUSSION.

6. RECOMMENDATIONS.

The last three sections will be discussed in detail in **Module 31**. The first three sections can be revised from the relevant sections in your research proposal.

Presentation of Summary of Findings and Recommendations

Since an important goal of your research is that appropriate action will be undertaken based on the results of your study, it is important that all parties concerned get an opportunity to discuss findings and recommendations before the report is finalized. You may wish to include policymakers, health managers, staff, and community members in such a discussion. **Module 32** provides some directions on how to organize a meeting for this purpose.

Drafting a Plan for the Implementation of the Research Results

You drafted a plan for utilization and dissemination of results during the previous workshop (**Module 16**). At the end of the present workshop this plan will be reviewed and developed in more detail.

GROUP WORK (time flexible, depending on the research topics and state of data processing)

- Reconsider the objectives from your research proposal and list the different data sources for each objective (questionnaires, records, focus group discussions etc.).

  NB If you discover that you have collected more data to explain your research problem than your objectives require, you may review your objectives or add one or two additional ones. However, if you collected less data, don't drop objectives which you could not meet but explain why you couldn't.

- Verify whether all data has been checked for completeness, consistency, and proper coding. If not, do so.

- Determine whether different mastersheets have been prepared for different study populations or for different categories of informants (one for defaulters and one for compliers, for example). This will facilitate analysis and, if required, comparison of groups.

  You may also mark the questionnaires of sub-groups with different colours so that you can easily refer back to the raw data to check certain questions.

- Check whether the mastersheets have been completed and whether the number of responses for each variable agree with the number of respondents.

  Determine whether all data that should have been entered in the computer have been entered and cleaned. (Refer to Annex 21.1.)

- Check whether qualitative data were categorized and summarized in the field. Record all your relevant observations.

Annex 21.1

COMPUTER OUTPUT

What is printed out (hard copy) by the computer is the result of the commands used in the programs to analyze the available data. The accuracy of the information printed out is therefore dependent on:

- the data that were entered,
- the programs that were used.

The saying, "garbage in, garbage out," is very apt for computer processing. It is the responsibility of the investigator to ensure that the information printed out is accurate.

Types of Computer Printouts

1. Lists of Data

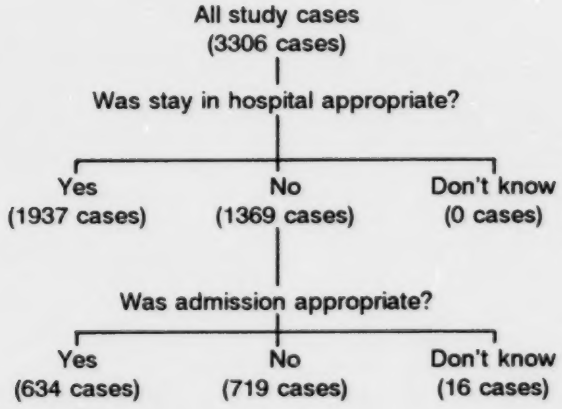
This is a list of the data that were entered into the computer. This printout is helpful if you need to make corrections to the existing data while in the process of validating it.

2. Frequency Count

This gives a count (and percentage) of each variable in the questionnaire. See the given sample of a frequency count and note how it related to the questionnaires.

To ensure that the programs are correct, the investigator must be familiar with the format or questionnaire used and the process of data collection. Developing a flow-chart of the process with pertinent questions and counts of the variables incorporated would help the investigator to check the accuracy of the programs.

For example:



In this example, the total responses for "yes", "no", and "don't know" for the question "Was the stay in hospital appropriate?" must equal the total study cases. If they do not, communicate with your computer specialist or statistician.

A frequency count should be obtained for every question in the questionnaire. Use the frequency count to ensure that:

- the total number of responses in each question is correct (i.e., it should tally with the sample size, or number of persons being asked the question);
- all codes are relevant to the question. **For example**, there should be no codes 3-8 in a question which has only 2 possible responses and an "unknown" (if unknown is given a code 9).

3. Cross-Tabulation

The next commonest computer output is a **cross-tabulation**. This is a table showing the number of subjects who have two (or more) of the variables studied.

Example:

	Male	Female
Ill		
Not ill		

Before using it, check the cross-tabulation for the following:

- The grand total in the table should correspond to the number of subjects in the sample;
- Column and row totals should correspond to the frequency counts for each variable (i.e., the number of males and females should correspond to the respective frequency counts);
- Similarly, numbers "ill" and "not ill" should correspond to that frequency count. If these do not correspond, there is probably an error in the program. Consult your computer specialist.
- If there is a statement in the computer printout showing "missing cases" it means either:
  - there is a wrong code in the data entry (e.g. code 4 when only 1,2 or 9 is possible), or
  - the categories you have specified are not comprehensive.

  For example:
  The questionnaire allowed for "unknown" but the computer programme did not. Therefore all cases "unknown" would appear as "missing cases."

  Marital status in the questionnaire allowed for "married, single, divorced, widowed." However, the computer program specified only "married, single, divorced." All widowed persons would be missing.

  If the age categories are 10 to 14, 15 to 19 but the programmer accidently programmed the categories as 10 to 13, 15 to 19, all subjects aged 14 would be missing.

Trainer's Notes

Module 21: ORIENTATION TO THE WORKSHOP ON DATA ANALYSIS AND REPORT WRITING

Timing and Teaching Methods	
3 hours	Presentations of field experiences
1/2 hour	Introduction and discussion
1 hour*	Group work (duration depending on research topic and state of data processing)

Introduction and Discussion

- Spend the first part of the introductory session on the participants' reports of their field experiences. If all the groups have come prepared to present their preliminary reports, the session can begin directly with this activity. However, if groups still need some time to prepare for their presentations, some time should be arranged for this, either before or at the beginning of this first session.

- The introduction to the workshop should clearly stress that there are different tasks to complete, of which data analysis and reporting of the findings will be the most time consuming. It should be clear to the participants, however, that the preparation of recommendations and implementation is the ultimate aim of their research projects. You might ask the participants for suggestions on which policymakers and managers should be invited for the presentation and discussion of their research findings and recommendations at the end of the workshop.

- When presenting the diagram, consider using different, overlapping transparency sheets.

- Adjust the presentation to the level and interests of the participants. Refresh their memory with examples of the processing of open-ended questions and by explaining the difference between descriptive studies, comparative studies looking for differences between groups, and studies determining associations between variables — **preferably with examples from their own research.**

- Do not frighten groups that have little statistical experience with details on tests at this stage. Merely state that each type of study requires different tests.

- Stress the importance of listing all the data available for each objective, including qualitative data. As the workshop proceeds there will be so much emphasis on the preparation of tables that participants will tend to forget valuable observations and information obtained from key informants. The facilitator should ask the participants to record this information now (if it is not already done) and include it in the list of data available for each objective. Check, when the report is being written, that it has been analyzed.

Group Work

- Read the group work directions along with the group members. Let them reexamine their objectives, list the data available for each objective, and discuss whether the objectives are specific enough to cover all relevant data collected. Sometimes objectives have to be split up, rephrased, added, or their order changed to facilitate analysis. **Never** allow the group to omit an objective without an explanation (in the methodology section) concerning why it could not be met.

- Examine, with the group members, all available data for completeness, mistakes, etc. Make sure that separate master sheets have been prepared for different study populations or for different subgroups that will be compared, or that the data on different subgroups can be easily retrieved from the computer.

- **Take extra time, as a facilitator, to internalize all data available, to identify possible weaknesses, and to consider various possibilities for analysis.** Unless you do this at the onset of the workshop it will be difficult to guide the groups efficiently so that they will obtain optimal results from the data they have collected.

- Group members may work in subgroups to finalize the data processing, but make sure that you discuss problems and progress with the group as a whole at regular intervals.

Health Systems Research Training Series

Volume 2, Part II: Data Analysis and Report Writing

Module 22:

DESCRIPTION OF VARIABLES: PART I

STEPS IN DATA ANALYSIS AND REPORT WRITING

Questions you must ask	Steps you will take*	Important elements of each step
What data have been collected for each research objective? Are data complete, accurate?	Prepare data for analysis	Review field experiences Inventory data for each objective/study population Sort data and check quality Check computer outputs
What do the data look like?	Describe variables	Frequency distributions Figures, means
How can the data be summarized for easy analysis?	Cross-tabulate quantitative data Summarize qualitative data	Cross-tabulate in relation to objectives Graphic displays, narratives
For quantitative data: does each research objective aim to describe, compare, or find associations?	Determine the type of statistical analysis required	Review objectives, study type, and variables Statistical description of variables Choosing significance tests
1. How can the data be described?		
2. How can differences between groups be determined?	Analyze paired and unpaired observations	**Student's t-test **Paired t-test **Chi-square test **McNemar's chi-square test
3. How can the associations between variables be determined?	Measure associations between variables	**Scatter diagram **Regression line and correlation coefficient **Relative risk, odds ratio
How should the report be written?	Write the report and formulate recommendations	Prepare outline for report Draft and redraft Summarize findings Summarize conclusions for each objective Formulate recommendations Prepare abstract
How should the findings and recommendations be presented and disseminated?	Present a summary and draft a plan for implementation of recommendations	Discuss summaries with different target groups Discuss plan for implementation

* These steps need not be in the sequence in this diagram. The sequence may be adjusted according to the needs of the workshop.
** These elements are optional and may be omitted if not relevant in a particular course.

Module 22: DESCRIPTION OF VARIABLES, PART I

OBJECTIVES

At the end of this session you should be able to:

1. Describe data in terms of frequency distributions, percentages, and proportions.
2. Use figures to present data.
3. Explain the difference between mean, median, and mode.
4. Calculate the frequencies, percentages, proportions, ratios, rates, means, medians, and modes for the major variables in your study that require such calculations.
5. Identify other independent variables (in addition to the ones identified during the first workshop), if any, that are necessary in the analysis of your data.

I. **Introduction**

II. **Frequency Distributions**

III. **Percentages, Proportions, Ratios, and Rates**

IV. **Figures**

V. **Measures of Central Tendency**

I. INTRODUCTION

When you selected the variables for your study in **Module 8**, you did so in the belief that they either helped to describe your problem (dependent variables) or that they were contributory factors to your problem (independent variables). The purpose of data analysis is to determine which variables best describe the problem and the factors influencing the problem, and how the data answer the research questions outlined in the objectives.

Before we look at how variables may be affecting one another, we need to summarize the information obtained on each variable in simple tabular form or in a figure.

Some of the variables may have produced numerical data, while other variables may be categorical and have produced data in categories of one type or another. In analyzing our data, it is important first of all to determine the type of data that we are dealing with. This is crucial in organizing our approach to statistical methods because the type of data used largely determines the general type of statistical techniques that are applicable.

Categorical Data

There are two types of categorical data: nominal and ordinal data.

In NOMINAL DATA, the variables are divided into a number of named categories. These categories however, cannot be ordered one above another (as they are not greater or less than each other).

For example:

NOMINAL DATA	CATEGORIES
Sex	male, female
Marital status	single, married, widowed, separated, divorced

In ORDINAL DATA, the variables are also divided into a number of categories, but they can be ordered one above another, from lowest to highest or vice versa.

For example:

ORDINAL DATA	CATEGORIES
Level of knowledge	good, average, poor
Opinion on a statement	fully agree, agree, disagree, totally disagree

Numerical Data

We speak of NUMERICAL DATA if they are expressed in numbers.

Some numerical variables are essentially **discrete**, such as numbers of motor accidents or prices of eggs. The possible values take only a distinct series of numbers. Actually all numerical data are discrete as recorded, since we round for simplicity.

For example:

- height to the nearest centimetre or inch,
- temperature in degrees Celsius,
- age to the last birthday.

But with such data, we can imagine the development of more and more accurate measuring instruments and greater detail in recording so that the possible recorded values increase without limit and the data become essentially continuous. In statistical writings such data are usually referred to as **continuous**.

Numerical data can be examined through:

- Frequency distributions,
- Percentages, proportions, ratios, and rates,
- Figures,
- Measures of central tendency.

We will now discuss these operations one after each other for both categorical and numerical data.

II. FREQUENCY DISTRIBUTIONS

A FREQUENCY DISTRIBUTION is a description of data presented in tabular form so the data will be more manageable. It gives the frequency with which (or the number of times) a particular value appears in the data.

In your research project you will have already done straight frequency counts for all variables on the basis of the data master sheets by counting the number of responses in each category. We will now briefly summarize some important points.

- **CATEGORICAL data** may have very simple categories.

  Example 1:

  To check the accuracy of the clinical diagnosis of malaria, blood slides from 33 patients were examined for malaria parasites. There were three possible results: Negative, *P. falciparum* or *P. vivax*.

The results are presented in the following **frequency distribution**:

Negative	19
P. falciparum	13
P. vivax	1
TOTAL	33

These data are **NOMINAL**. A **frequency distribution** is calculated by simply totalling the number of responses in each category.

You should always check that the total number of responses agrees or tallies with the number of subjects (respondents). If necessary there should be a category for missing answers.

By looking at the frequency distribution you can conclude that more than half of the patients with a clinical diagnosis of malaria actually do not have malaria. Another observation would be that *P. falciparum* is much more common than *P. vivax*.

Example 2:

Health personnel from 148 different rural health institutions were asked the following question: "How often have you run out of drugs for the treatment of malaria in the past two years?"

This was a closed question with the following possible answers: never, 1 to 2 times (rarely), 3 to 5 times (occasionally), more than 5 times (frequently). The numbers of responses in each category were totalled to give the following frequency distribution.

Never	47
Rarely	71
Occasionally	24
Frequently	6
TOTAL	148

In this example, the data are **ORDINAL**. The ordering of the categories is important as each category from top to bottom indicates increasing severity of the problem.

The frequency distribution indicates that most clinics either do not or rarely experience shortages of antimalarial drugs, but that it is an occasional problem in about a sixth of the clinics and a severe problem in a few.

- Procedures for making frequency distributions of **NUMERICAL** data are very similar to those for categorical data, except that now the data have to be grouped in categories. The steps involved in making a frequency distribution are as follows:

1. Select groups for grouping the data.
2. Count the number of measurements in each group.
3. Add up and check the results.

When grouping data, the way the groups are selected can affect what the results are going to look like. There is little substitute for common sense here, but it may be necessary to change the grouping if you suspect the information is being hidden by a poor selection of the groups.

Example 3:

Numbers of malaria cases are being submitted daily by health centres of District X and you wish to summarize them. Compare the following daily and weekly summaries of the same data:

Table 22.1. Daily and weekly summaries of malaria cases in health centres in District X.

Day 1	9 cases		
Day 2	12		
Day 3	11		
Day 4	13		
Day 5	14		
Day 6	13		
Day 7	16	Week 1	88 cases
Day 8	16 cases		
Day 9	16		
Day 10	18		
Day 11	19		
Day 12	16		
Day 13	21		
Day 14	25	Week 2	131 cases
Day 15	28		
Day 16	28		
Day 17	28		
Day 18	32		
Day 19	21		
Day 20	19		
Day 21	12	Week 3	168 cases

Both the daily and the weekly summaries show an increasing number of cases of malaria. However, if we present only the weekly data the improving situation shown on days 19, 20, and 21 will not be reflected. It would, therefore, be better to present the daily data, if you want to show exactly when the number of reported malaria cases is going down.

When grouping data the following rules are important:

- The groups must not overlap, otherwise there is confusion concerning in which group a measurement belongs.
- There must be a continuity from one group to the next, which means that there must be no gaps. Otherwise some measurements may not fit in a group.
- The groups must range from the lowest measurement to the highest measurement so that all of the measurements have a group to which they can be assigned.
- The groups should normally be of an equal width, so that the counts in different groups can easily be compared.

Sometimes, however, it is valid to choose groups which are of different widths, for example if you are interested in specific age groups (e.g., less than 1 year, 1 to 4 years, 5 to 14 years).

When you start summarizing data it is better to make too many groups than too few. This is because you can combine groups to form new categories without having to go through all your data again, whereas if you have too few groups you have to go through the data again to make more groups.

A larger number of groups will generally give a more precise picture, but when using too many groups one can lose sight of the overview.

As a general rule choose round numbers for the lower values of the group limits.

For example: 1.00-9.99, 10.00-19.99, 20.00-29.99

III. PERCENTAGES, PROPORTIONS, RATIOS, AND RATES

Percentages

Instead of presenting data in frequency tables using absolute numbers it is often better to calculate percentages.

> A PERCENTAGE is the number of units with a certain characteristic divided by the total number of units in the sample and multiplied by 100.

Percentages may also be called RELATIVE FREQUENCIES. Percentages standardize the data, which means that they make them easier to compare with similar data obtained in another sample of a different size.

Example 4:

82 Clinics in one district were asked to submit the number of patients treated for malaria in one month. The researchers presented both the frequency distribution and percentages (or relative frequencies):

Table 22.2. Distribution of clinics according to number of patients treated for malaria in one month.

Numbers of patients	Number of clinics[a]	Relative frequency
0 to 19	25	31%
20 to 39	3	4%
40 to 59	5	6%
60 to 70	11	14%
80 to 99	19	24%
100 to 119	10	12%
120 to 139	4	5%
140 to 159	3	4%
Total	80	100%

[a] Data from two clinics are missing.

Note:
Usually you do not include missing data in the calculation of percentages.

The frequency of responses in each group is calculated as the percentage of those study elements for which you obtained the data (or if a question is being asked to interviewees, the percentage of those interviewees who answered the question).

However, the number of missing data (e.g., people who did not respond to a question) is a useful indication of the adequacy of your data collection. Therefore, this number should be mentioned, for example as a note to your table. (See **Table 22.2.**)

Remember that "don't know" is a category on its own and should not be counted as missing data. If applicable, "don't know" should appear as a category in the table.

Caution should be applied when calculating and interpreting percentages when the total number is small, because one unit more or less would make a big difference in terms of percentages.

For instance, in **Example 1** the total number of blood slides examined is only 33. P falciparum is found in 13 blood slides, which corresponds to 39%. If P. falciparum had been found 14 instead of 13 times (which easily could have been the case), then the relative frequency would have been 42%, which already makes a difference of 3%.

If the total number of blood slides examined had been 330, out of which 130 (or 39%) had P. falciparum, then one slide more or less (131 or 129) would not have made such a big difference in terms of percentages.

Therefore it is recommended that the number of observations or total cases studied be given together with the percentage.

Proportions

Sometimes relative frequencies are expressed in proportions instead of percentages.

A PROPORTION is a numerical expression that compares one part of the study units to the whole;
A proportion can be expressed as a FRACTION or in DECIMALS.

Example 5:

Out of a total of 55 patients attending a clinic on a specific day 22 are males and 33 are females. We may say that the proportion of males is 22/55 or 2/5, which is equivalent to 0.40. (The numerator is 22, the denominator is 55).

Note that when a proportion expressed in decimals is multiplied by 100 the value obtained is a percentage. In the example 0.40 is equivalent to 40%.

Ratios

A **RATIO** is a numerical expression which indicates the relationship in quantity, amount or size between two or more parts.

In **Example 5** the ratio of males to females is 22:33, which is 2:3.

Rates

A **RATE** is the quantity, amount or degree of something measured in a specified period of time.

Commonly used rates in the health sector are:

- Birth Rate = The number of live births per 1000 population over a period of one year.
- Death Rate = The number of deaths per 1000 population over a period of one year.
- Infant Mortality Rate = The number of infant deaths per 1000 live births
- Maternal Mortality Rate = The number of maternal pregnancy-related deaths in one year per 100,000 total births in the same year

IV. FIGURES

If your report contains many descriptive tables, it may gain in readability if you present the most important ones in figures.

The most frequently used figures for presenting data include:

- Bar charts
- Pie charts

  → for categorical data

- Histograms
- Line graphs
- Scatter diagrams

  → for numerical data

- Maps

We will now look at examples of the abovementioned figures that can be used for presenting data.

Bar Chart

The data from **Example 2** can be presented in a bar chart, using either absolute frequencies (see **Figure 22.1**) or relative frequencies or percentages (see **Figure 22.2**).

Figure 22.1. Frequency of shortage of antimalaria drugs in rural health institutions.

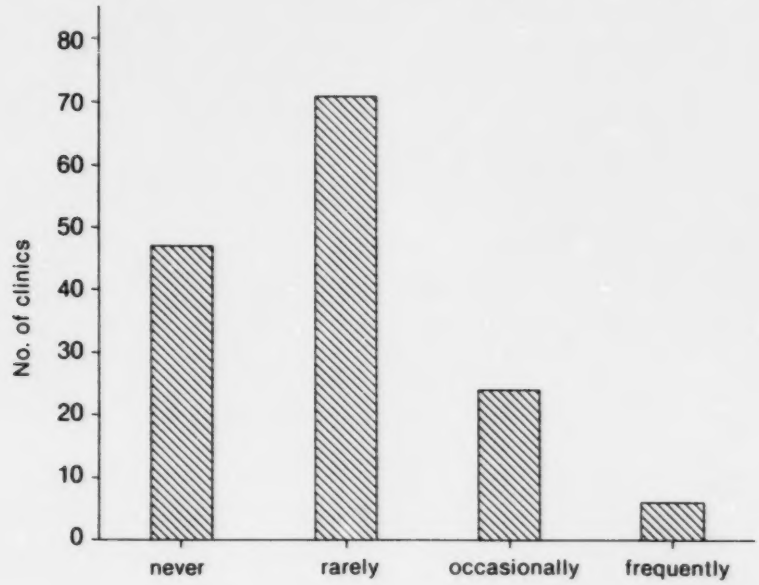
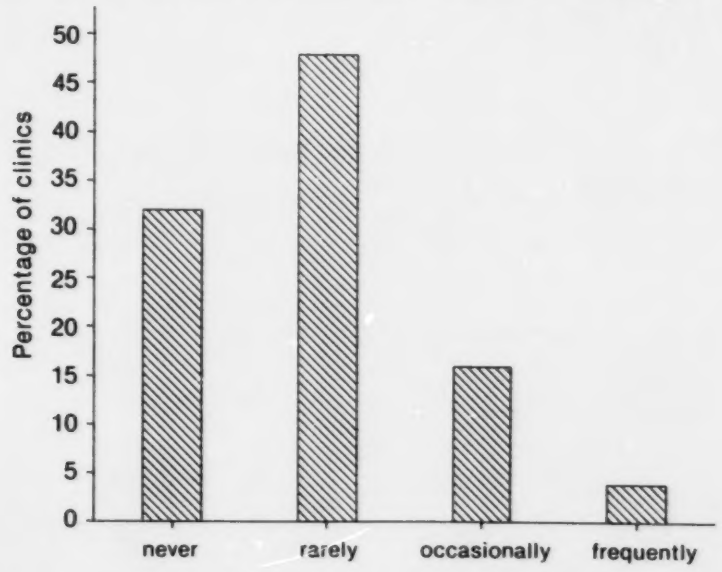


Figure 22.2. Relative frequency of shortage of antimalaria drugs in rural health institutions (n = 148).

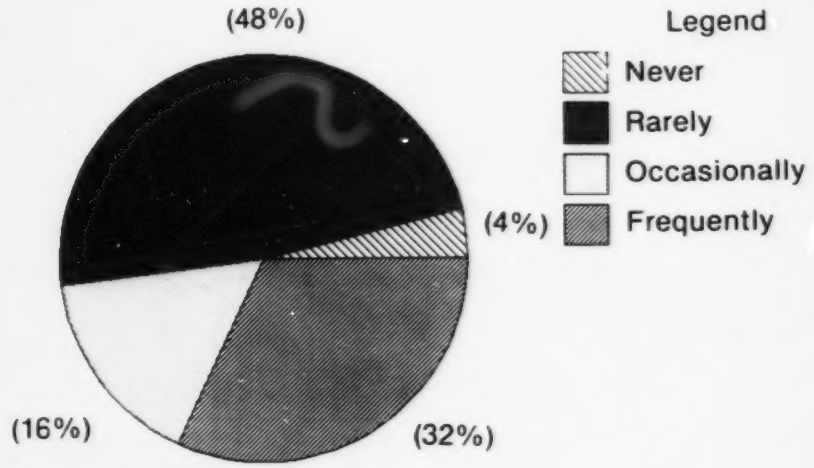


Note that the sample size must be indicated if you present the data in percentages.

Pie Charts

A pie chart can be used for the same set of data, providing the reader with a quick overview of the data presented in a different form. A pie chart illustrates the relative frequency of a number of items, i.e., all the segments of the pie chart should add to 100%

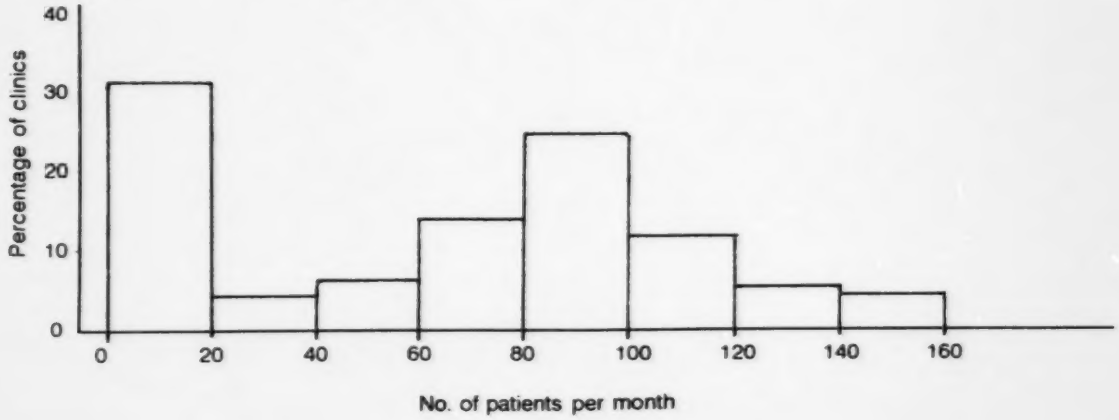
Figure 22.3. Relative frequency of shortage of antimalaria drugs in rural health institutions (n = 148).



Histograms

Numerical data are often presented in histograms, which are very similar to the bar charts that are used for categorical data. An important difference, however, is that in a histogram the "bars" are connected (as long as there is no gap between the data), whereas in a bar chart the bars are not connected, as the different categories are distinct entities. The data of **Example 4** are presented as a histogram in **Figure 22.4.**

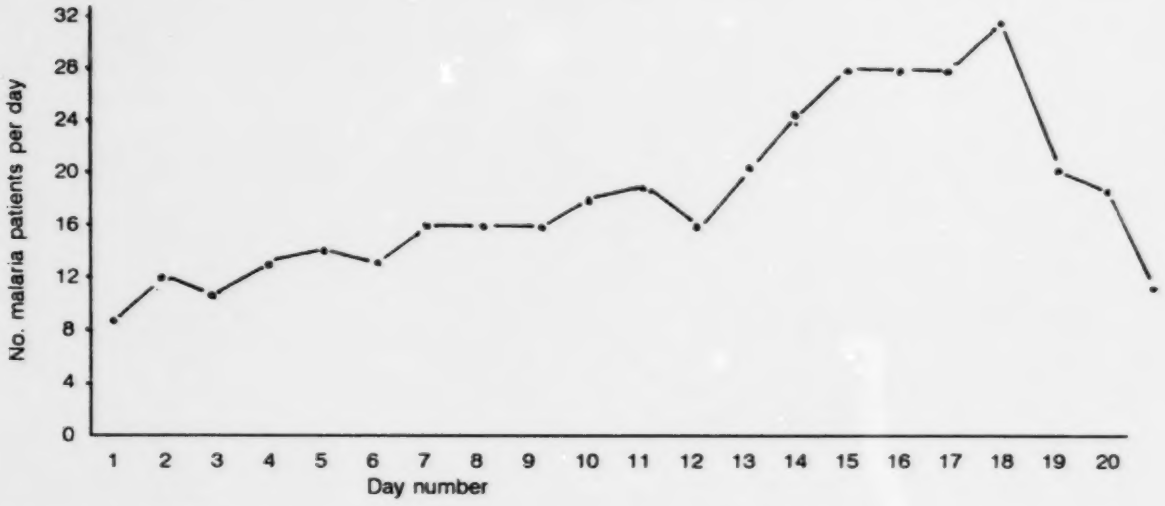
Figure 22.4. Percentage of clinics treating different numbers of malaria patients (n = 80).



Line Graphs

A line graph is particularly useful for numerical data if you wish to show a trend over time. The data from **Example 3** can be presented as in **Figure 22.5.**

Figure 22.5. Number of malaria patients per day at the health centres in District X.

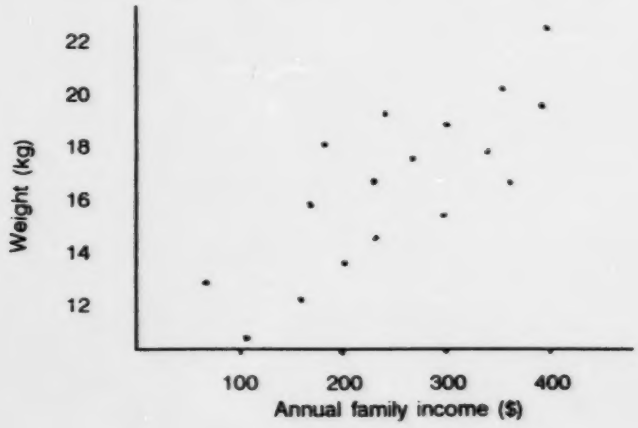


It is easy to show two or more distributions on one graph, as long as the lines are easy to distinguish. Thus it is possible to compare frequency distributions.

Scatter Diagrams

Scatter diagrams are useful for showing information on two variables which are possibly related. The example of a scatter diagram given below is used in **Module 29**, where we are dealing with the concepts of association and correlation.

Figure 22.6. Weight of 5 year olds according to annual family income.



Note:

It is important that all figures presented in your research report carry a number, a clear title and clear labels (or keys).

In addition to the figures above, the use of **MAPS** may be considered to present information. For instance, the area where a study was carried out can be shown in a map. If the study explored the epidemiology of cholera, a map could be produced showing the geographical distribution of cholera cases, together with the distribution of protected water sources, thus illustrating that there is an association. If the study related to vaccination coverage, a map could be developed to indicate the clinic sites and the vaccination coverage among under-fives in each village, perhaps showing that home–clinic distance is an important factor associated with vaccination status.

V. MEASURES OF CENTRAL TENDENCY

Frequency distributions and histograms provide useful ways of looking at a set of observations of a variable. In many circumstances, it is essential to produce them to understand the patterns in the data. However, if one wants to further summarize a set of observations, it is often helpful to use a measure that can be expressed in a single number.

First, one would like to have a measure for the centre of the distribution. The three measures that are used for this are the **MEAN**, the **MEDIAN**, and the **MODE**.

Mean

The MEAN (or arithmetic mean) is also known as the AVERAGE. It is calculated by totalling the results of all the observations and dividing by the total number of observations. Note that the mean can only be calculated for numerical data.

Example 6:

Measurement of the height of 7 women gave the following results:

141, 141, 143, 144, 145, 146, 155 cm (a total of 1015 cm for 7 measurements)

The mean is thus 1015/7, which is 145 cm.

Median

The MEDIAN is the value that divides a distribution into two equal halves.

The median is useful when some measurements are much bigger or much smaller than the rest. The mean of such data will be biased toward these extreme values. Thus the mean is not a good measure of the centre of the distribution in this case. The median is not influenced by extreme values.

The median value, also called the central or halfway value, is obtained in the following way:

- List the observations in order of magnitude (from the lowest to the highest value).
- Count the number of observations (n).
- The median value is the value belonging to observation number (n + 1) / 2.

Example 8:

The weights of 7 pregnant women are:

40, 41, 42, 43, 44, 47, and 72 kg.

The median value is the value belonging to observation number (7 + 1) / 2 = the fourth one, which is 43 kg.

Note that the mean weight of this set of observations is 47 kg. This is an illustration of how the mean is affected by extreme values (in this case 72 kg) while the median isn't. If the largest weight in this set of observations had been 51 kg instead of 72 kg, the median would still have been 43 kg, but the mean weight would have been 44 kg.

Mode

The MODE is the most frequently occurring value in a set of observations.

The mode is not very useful for numerical data that are continuous. It is most useful for numerical data that have been grouped into classes.

In **Example 4** (number of patients treated for malaria at clinics) the mode is "0 to 19," as this outcome is recorded most frequently (25 times out of 80).

The mode can also be used for categorical data, whether they are nominal or ordinal.

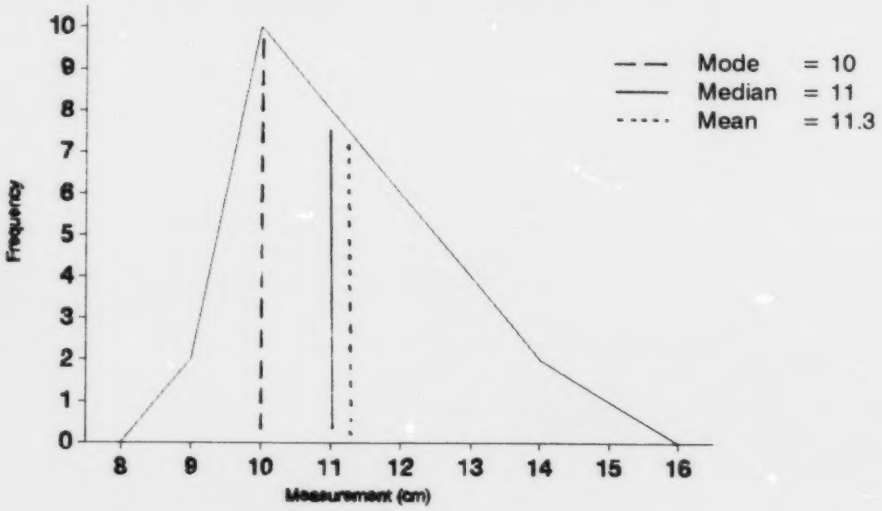
In **Example 1** (clinical diagnosis of malaria) the mode is "negative." In **Example 2** (number of clinics experiencing drug shortage) the mode is "rarely."

In summary, the mean, the median, and the mode are all measures of central tendency. The mean is most widely used. It contains more information because the value of each observation is taken into account in its calculation.

However, the mean is strongly affected by values far from the centre of the distribution, while the median and the mode are not. The calculation of the mean forms the beginning of more complex statistical procedures to describe and analyze data.

Figure 22.7 shows a distribution curve in which the mean, the median and the mode have different values.

Figure 22.7. Mean, median, and mode in a distribution curve.



GROUP WORK

- Describe your sample(s) in terms of background variables (sex, age, etc.) and dependent variables (e.g., defaulter/complier, user/non-user).
- Make sure that you have made frequency counts for all variables in your study (from your data master sheets). Calculate percentages in relation to the total number of study units (or calculate proportions/ratios/rates where appropriate).
- Check your objectives to determine which variables require frequency tables that should be included in your report. Usually frequency tables are presented for some of the background variables, the dependent variable(s), and the most important independent variables. Prepare the frequency tables.
- Make histograms, bar charts, pie charts, and line graphs, if useful. Prepare brief descriptions that interpret what each of the figures mean.
- Calculate means, medians, and modes, if applicable, and interpret the results.
- Familiarize yourself with the results and try to understand as fully as possible what they mean.

Trainer's Notes

Module 22: DESCRIPTION OF VARIABLES, PART I

Timing and Teaching Methods	
1 hour	Introduction and discussion
3 hours*	Group work

Introduction and Discussion

- It is likely that the participants will be familiar with some of the concepts introduced in this module, such as percentages and proportions. Moreover, at this stage, the groups will have already prepared frequency distributions (including calculation of percentages). Therefore these concepts should be only briefly mentioned in the presentation, especially if the knowledge level of the participants is high, so as not to lose their interest. However, special attention should be given to what to do with missing values when calculating percentages.

- Although definitions of percentage, proportion, ratio, and rate are given in the module, it is more important to provide examples or ask participants to provide their own.

- When presenting **Example 3**, you might also ask participants to describe how they have grouped numerical data and discuss whether there were too few or too many categories.

- Examples should not be merely presented, they should be used in informal exercises. For example, ask participants what is the mean, median, and mode of a given set of measurements instead of providing them with the answers.

Group Work

- Before the group makes frequency counts for all variables from the master sheets, have them check whether the data have been categorized correctly. Also be sure that the total number of informants (study units) for all groups that have been studied has been defined.

- Remind the group that fully developed frequency tables are only necessary for variables that have to be described in the final report. Usually tables are needed for some of the background variables, the dependent variable(s), and sometimes for the most important independent variables. Many of the other background and independent variables will be presented using cross-tabulations (**Module 23**).

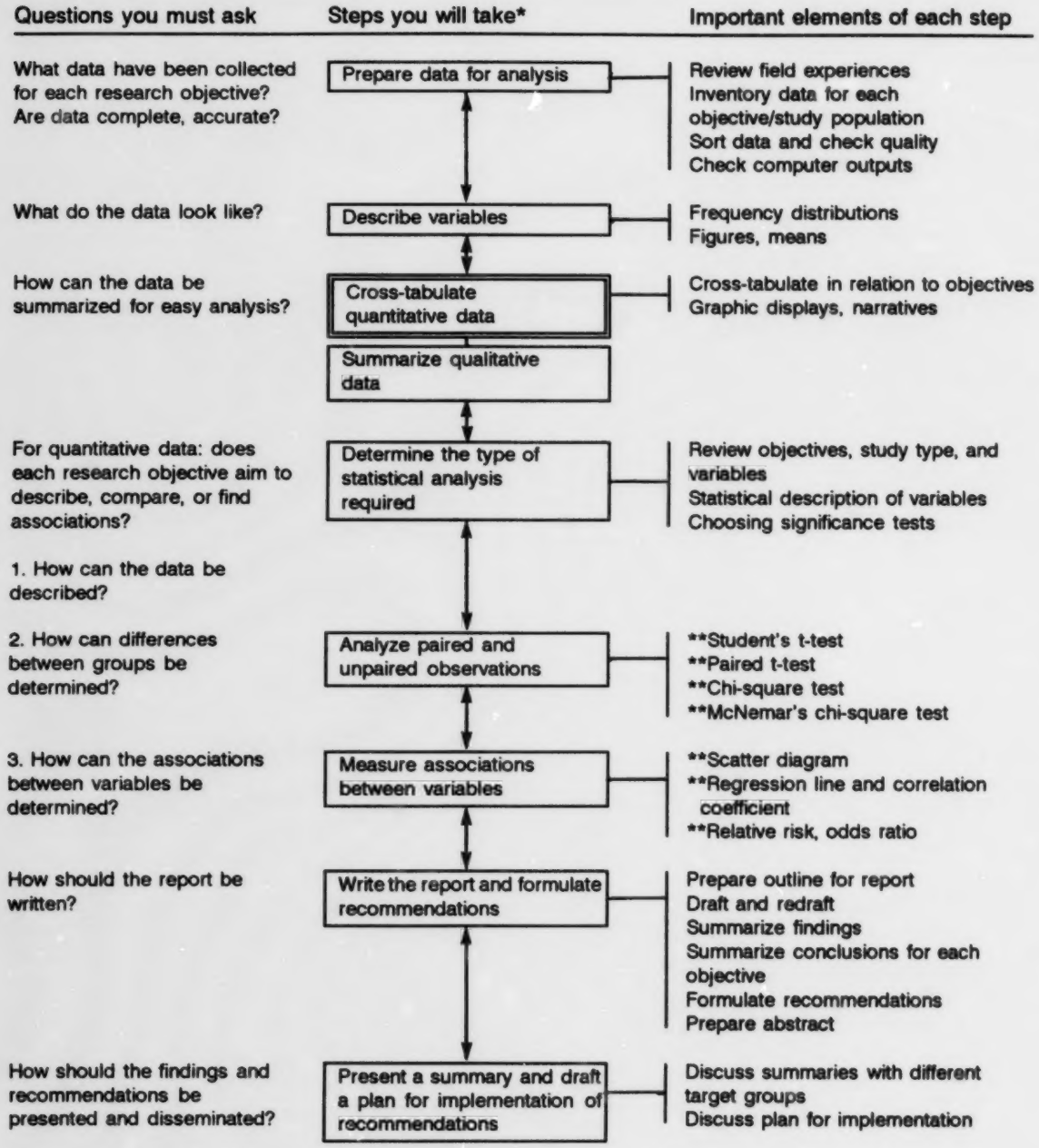
Health Systems Research Training Series

Volume 2, Part II: Data Analysis and Report Writing

Module 23:

CROSS-TABULATION OF QUANTITATIVE DATA

STEPS IN DATA ANALYSIS AND REPORT WRITING


Questions you must ask
Steps you will take*
Important elements of each step
What data have been collected for each research objective? Are data complete, accurate?
Prepare data for analysis
Review field experiences
Inventory data for each objective/study population
Sort data and check quality
Check computer outputs
What do the data look like?
Describe variables
Frequency distributions
Figures, means
How can the data be summarized for easy analysis?
Cross-tabulate quantitative data
Summarize qualitative data
Cross-tabulate in relation to objectives
Graphic displays, narratives
For quantitative data: does each research objective aim to describe, compare, or find associations?
Determine the type of statistical analysis required
Review objectives, study type, and variables
Statistical description of variables
Choosing significance tests
1. How can the data be described?
2. How can differences between groups be determined?
Analyze paired and unpaired observations
**Student's t-test
**Paired t-test
**Chi-square test
**McNemar's chi-square test
3. How can the associations between variables be determined?
Measure associations between variables
**Scatter diagram
**Regression line and correlation coefficient
**Relative risk, odds ratio
How should the report be written?
Write the report and formulate recommendations
Prepare outline for report
Draft and redraft
Summarize findings
Summarize conclusions for each objective
Formulate recommendations
Prepare abstract
How should the findings and recommendations be presented and disseminated?
Present a summary and draft a plan for implementation of recommendations
Discuss summaries with different target groups
Discuss plan for implementation

* These steps need not be in the sequence in this diagram. The sequence may be adjusted according to the needs of the workshop.
** These elements are optional and may be omitted if not relevant in a particular course.

Module 23: CROSS-TABULATION OF QUANTITATIVE DATA

OBJECTIVES

At the end of this session you should be able to:

1. **Construct** all important cross-tabulations that will help meet your research objectives.
2. **Identify** possible confounding variables that need consideration when exploring relationships between variables, and take appropriate measures to deal with them.

I. **Introduction**

II. **Different Types of Cross-Tabulations**

III. **Constructing Cross-Tabulations Appropriate for the Research Objectives**

IV. **General Hints When Constructing Tables**

V. **Dealing with Confounding Variables: Stratification and Matching**

I. INTRODUCTION

Thus far we have made tables containing frequency distributions for **one variable at a time** to partially describe our data. Depending on the **objectives** of our study and the **study type**, we may have to **examine the relationship between several of our variables** at once in order to adequately describe our problem or identify possible explanations for it.

For this purpose it is appropriate to design CROSS-TABULATIONS.

Example 1:

We want to know the ages at which teenage pregnancies occur and whether they are more frequent among school girls than among girls who are not attending school. To answer these questions we may construct the following cross-tabulation. (The data are imaginary.)

Table 23.1. Number of teenage pregnancies at different ages among girls attending school and not attending school (Province X, 1988-1990).

Age at onset of pregnancy	Number of pregnancies		Total
	Attending school	Not attending school	
12 years	2 (3%)	1 (2%)	3
13 years	2 (3%)	0	2
14 years	5 (7%)	2 (4%)	7
15 years	23 (34%)	12 (23%)	35
16 years	36 (53%)	37 (71%)	73
Total	68 (100%)	52 (100%)	120

Cross-tabulation tables, such as Table 23.1, that describe a problem or a situation, are called DESCRIPTIVE TABLES. Cross-tabulation tables that aim to discover possible explanations for a problem or that describe the results of an intervention are called ANALYTIC TABLES.

Some examples of descriptive and analytic cross-tabulation tables are provided in the following pages. You will notice that in each example the nature and the design of the cross-tabulation table is determined by the objectives and the type of study conducted.

II. DIFFERENT TYPES OF CROSS-TABULATIONS

Depending on the objectives and the type of study, three different kinds of cross-tabulations may be required:

- **Cross-tabulations to describe the sample** aim at describing the problem under study by presenting a combination of variables. **Table 23.1** is an example. Descriptive cross-tabulations are also used to describe the sample of research subjects in terms of a combination of background variables, such as age, sex, profession, residence.

- **Cross-tabulations in which groups are compared to determine differences.**

- **Cross-tabulations that focus on exploring relationships between variables.**

Examples of each of these three types of cross-tabulations are given below. Note that tables that are not yet filled in with data are called DUMMY TABLES.

Cross-Tabulations to Describe the Sample

In any study which yields quantitative data, whether descriptive or analytic, it is common to first describe the research subjects included in the sample(s) before presenting the actual results of the study. This can be done for separate variables in a simple frequency table (as shown in **Module 22**) or for a combination of variables in a cross-table.

Example 2:

A study was carried out on the degree of job satisfaction (dependent variable) among doctors and nurses in rural and urban areas. To describe the sample a cross-tabulation was constructed which included the sex and the residence (rural or urban) of the doctors and nurses interviewed. This was useful because in the analysis the opinions of male and female staff had to be compared separately for rural and urban areas.

Table 23.2. Residence and sex of doctors and nurses.

Health workers		Residence		Total
		Rural	Urban	
Doctors	Males	8 (10%)	35 (21%)	43
	Females	2 (3%)	16 (10%)	18
Nurses	Males	46 (58%)	36 (22%)	82
	Females	23 (29%)	77 (47%)	100
Total		79 (100%)	164 (100%)	243

Table 23.2 can also be split up into two separate cross-tabulations for residence and type of health workers: one for males and one for females.

The data in the tables is usually listed in absolute figures as well as in relative frequencies (percentages or proportions). For numerical data (such as age) the mean, median and/or mode may be calculated as well to describe the sample.

In descriptive studies that aim at quantification of a certain problem, cross tabulation is a useful tool for presenting findings.

Example 3:

A study was constructed to examine the factors contributing to the high proportion of still-births in a hospital. The following dummy descriptive cross-tabulation would show how many of the fresh and macerated still-births weighed less than 2500 grams and how many weighed 2500 grams or more.

Table 23.3. Weight of fresh and macerated fetuses.

Weight of fetus	Condition at birth		Total
	Fresh	Macerated	
Less than 2500 g			
2500 g or more			
Total	(100%)	(100%)	

Note: The actual numbers would be placed in each cell of the table after data collection.

Cross-Tabulations to Determine Differences Between Groups

In comparative studies such as case-control studies, cohort studies or (quasi)-experimental studies, some objectives will focus on discovering whether any differences exist between two or more groups on particular variables. In such cases cross-tabulation of data is necessary.

Example 4:

In a quasi-experimental study of the effect of attendance of mothers at nutrition talks on their level of nutritional knowledge two groups of mothers were compared: those who did and those who did not attend the nutrition talks. The following dummy cross-tabulation table was prepared:

Table 23.4. Number of attenders and non-attenders at nutrition talks with different levels of nutritional knowledge.

	Level of nutritional knowledge			Total
	Low	Average	High	
Attenders				
Non-attenders				
Total				

In this example "low", "average" and "high" level of nutritional knowledge have to be clearly defined. (See **Module 8**).

Note that in the above cross-tabulation the groups that are to be compared are put in rows, whereas the different levels of nutritional knowledge are put in columns. This is because "nutritional knowledge" is the outcome of the attendance at nutrition talks and it is therefore considered to be the dependent variable. Dependent variables usually are displayed in columns.

Example 5:

In a case-control study on malnutrition, severely malnourished children (cases) were compared with well-nourished children (controls) in order to find appropriate ways of alleviating the problem of malnutrition in children. The following dummy cross-tabulation was constructed:

Table 23.5. Level of mothers' nutritional knowledge for mothers with severely malnourished and well-nourished children (12-35 months).

Level of mothers' nutritional knowledge	Nutritional status of their children		
	Severely malnourished (cases)	Well-nourished (controls)	Total
Low			
Average			
High			
Total	(100%)	(100%)	

Note the difference between Tables 23.4 and 23.5. In Table 23.5 "nutritional knowledge" is the independent variable, and thus is displayed in the rows, whereas the two groups which are to be compared are displayed in the columns.

Other analytic cross-tabulations can be constructed to be used in the study mentioned in Example 5. Each time the two groups (severely malnourished children and well-nourished children) can be displayed in the columns. Different independent variables can be put in rows, such as source of drinking water (protected or unprotected) or immunisation status (fully immunised or not).

Notes:

When constructing analytic cross-tabulations to detect differences between groups, carefully decide whether the groups that are to be compared should be put in rows or in columns, depending on the type of study. (The accepted convention is to place the categories of the dependent variable as column headings and the categories of the independent variables as row headings.)

When calculating percentages to be put in the cross-table it is important to remember that the totals for each of the groups which are to be compared should be 100%. (Look at the difference between Tables 24.4 and 24.5.)

Once these cross-tabulations are filled in with the appropriate data they will be used when performing statistical tests. These tests will be discussed in **Modules 27 and 28**.

Cross-Tabulations to Explore Relationships Between Variables

Many health systems research projects aim at exploring possible relationships or associations among variables. In such cases it is important to consider whether the variables are independent and dependent.

Example 6:

In a study of breast-feeding practices you may examine the relationship between the age of mothers and duration of breast-feeding. The cross-tabulation presented in Table 23.6 could be helpful.

Table 23.6. Duration of breast-feeding in mothers of different age groups.

Age (years)	Duration of breast-feeding			Total
	0-5 months	6-11 months	>12 months	
15-19				
20-24				
25-29				
30-34				
35-39				
40 +				
Total				

If you want to determine whether there is an association between the working status of mothers (independent) and the duration of breast-feeding (dependent variable) Table 23.7 would be appropriate.

Table 23.7. Duration of breast-feeding in relation to working status of mother.

Mothers' working status	Duration of breast-feeding			Total
	0-5 months	6-11 months	>12 months	
Full-time employed				
Part-time employed				
Not employed				
Total				

Analytic cross-tabulations that have been prepared to explore relationships between variables can be subjected to statistical analysis. **Modules 29 and 30** will explore this possibility further.

III. CONSTRUCTING CROSS-TABULATIONS APPROPRIATE FOR THE RESEARCH OBJECTIVES

When designing your research project you were asked to produce dummy tables for the data you expected to collect (**Module 13**). These dummy tables were made on the basis of the objectives and the type of study.

Since you now have collected your data and have an idea of their quality and how they can be used, you need to look again in a systematic way at the cross-tabulations to be made.

To construct appropriate cross-tabulations we recommend that you follow the steps below:

1. Review each specific objective and the method chosen for collecting the relevant data.

2. Formulate hypothetical sentences that you consider to be the type of conclusions you expect to reach concerning each objective.

   For example, in a descriptive study on breast feeding practices, where one of the specific objectives is "to determine factors associated with early weaning" expected conclusions could be:

   - "Mothers who are employed wean their children earlier than mothers who are not employed."
   - "Mothers who did not attend nutrition talks wean their children earlier than mothers who attended nutrition talks".

   The reasons for formulating "possible conclusions" are that they help you:

   - remember the purpose of each tabulation and calculation you undertake;
   - avoid wasting time on meaningless calculations and tabulations; and
   - keep your data organized so you can more easily write a well-organized report.

3. For each "expected conclusion" construct the dummy cross-tabulations that will enable you to derive the right conclusions.

4. Perform the appropriate frequency counts (using the data mastersheets) and enter the results in the cells of the cross-table.

5. Interpret the table and write a clear conclusion. It is not necessary to describe the content of the table in detail.

EXERCISE 1:

Select one specific objective from each of the research projects, formulate "expected conclusions" and construct the appropriate dummy cross-tabulation.

IV. GENERAL HINTS WHEN CONSTRUCTING TABLES

- Make sure that all the categories of the variables presented in the tables have been specified and that they are mutually exclusive (i.e. no overlaps and no gaps).
- When making cross-tabulations check that the column and row counts correspond to the frequency counts for each variable.
- Also check that the grand total in the table corresponds to the number of subjects in the sample.
- Think of a clear title for each table. Also be sure that the headings of rows and columns leave no room for misinterpretation.

- Number your tables and keep them together with the objectives to which they are related. This will assist in organizing your report and ensure that work is not duplicated.

V. DEALING WITH CONFOUNDING VARIABLES: STRATIFICATION AND MATCHING

Stratification

If one or more of the objectives of your study focus on exploring relationships among variables it is important to determine whether there are other factors which influence these relationships. These are known as CONFOUNDING VARIABLES. (See **Modules 8 and 9**.)

In **Example 6** it might be that the educational level of the mother is such a confounding variable, as it could be related to both the working status of the mother and duration of breast feeding. (See Table 23.7.)

Once you have collected your data the appropriate way of dealing with confounding variables is STRATIFICATION, which involves a SEPARATE analysis for the different levels of this confounding variable.

For **Example 6**, such a separate analysis involves the construction of cross-tabulations for mothers with different educational levels. This means that Table 23.7 has to be split up into two tables: one for those mothers who have little schooling, say less than 5 years, and another one for those who have 5 or more years of schooling.

If we find a SIMILAR association between working status and duration of breast feeding in both groups of mothers, then this indicates that the educational level of the mother is NOT a confounding variable.

Annex 3 of **Module 27** presents the statistical test to be used when dealing with confounding variables.

Matching

Module 9 (Study Type) stated that if at the stage of designing the study a variable is already suspected to be a confounding variable, the effect of this variable may be removed by PAIRING the observations (also called MATCHING). In this procedure each subject in the study group is matched with another subject in the so-called control group for the particular confounding variable.

In **Example 6**, for each employed mother with less than 5 years of schooling you would choose a non-employed mother with a similar educational level. Also for each employed mother with 5 or more years of schooling you would select a non-employed mother with 5 or more years of schooling.

It is important to take into account the fact that data are paired when constructing cross-tabulations and doing the analysis.

The analysis of paired observations will be examined in more detail in **Module 28** (where groups are compared in order to detect differences) and **Module 30** (Part 2 of section III: Case-control studies with paired observations).

GROUP WORK

- Review each specific objective and its relevant research design: formulate hypothetical sentences that describe the type of conclusions you expect for each objective.

- Construct dummy cross-tabulations, keeping in mind whether you want to:

  - describe research subjects in your sample or describe the problem;
  - compare groups in order to find differences; or
  - find associations between variables.

  Refer back to the dummy cross-tabulations which you already made in the first workshop (**Module 13**).

- If you construct analytic cross-tabulations, try to identify possible confounding variables by making further dummy cross-tabulations (stratification).

- Finally, fill in the dummy cross-tabulations with data and formulate a conclusion directly under each table.

Trainer's Notes

Module 23: CROSS-TABULATION OF QUANTITATIVE DATA

Timing and Teaching Methods	
1 hour	Introduction and discussion
3 hours+	Group work
1 hour	Group presentations and plenary discussion (optional)

Introduction and Discussion

- It is recommended that you use an overhead projector (or flipcharts) when presenting and explaining the construction of different tables in order to focus the attention of the participants. Do not merely refer to the modules.

- Some of the cross-tabulations presented in the module are filled in with imaginary data. This was done in order to make examples more concrete and to show how tables should be interpreted. However, special attention should be given to the design of the tables (what comes in rows, what comes in columns). Therefore it is recommended that you use 2 transparencies, which can be placed on top of each other, to present the table: one with the dummy table, the other with the data.

- Pay extra attention to how each table should be read: some of them have to be read horizontally, others vertically, depending on whether the groups to be compared are put in rows or in columns. Actually, the 100% which is already inserted in most of the tables helps to indicate how the table should be read.

- The classroom exercise at the end of section III helps to give participants useful practice in designing appropriate cross-tables. Try to obtain at least one table of each of the three types (i.e. descriptive and the two types of analytic tables) from the groups.

  Make sure that there is agreement concerning how to read the table (horizontally or vertically). If groups are having difficulty constructing appropriate tables for their own projects, this indicates that there is a need for the groups to present a few of their cross-tabulations in plenary after the group work session so they can get feed-back.

- The hints on constructing tables (section IV of the module) should be illustrated with an example of a table taken from the module or from one of the groups.

- Section V of the module (dealing with confounding variables) should be kept simple. All groups, however, should look for possible confounding variables so they will avoid coming up with meaningless conclusions.

Group Work

- Go through the first three steps of the group work assignment with the group as a whole. The data in the tables can be filled in by sub-groups.

- Stress that, once the tables are filled in with data, they should be interpreted immediately. The tentative conclusions should be recorded right under each table to make the writing of the report during the second week easier. Later these conclusions can be discussed with the other group members.

- At this stage the numbering of tables can be done according to the objective to which they relate.

Group Presentations and Plenary Discussion (Optional)

- Ask all groups to present at least one table (filled in with data) as well as the conclusions derived for each of their objectives which require cross-tabulations. Other groups and facilitators should be invited to comment.

Health Systems Research Training Series

Volume 2, Part II: Data Analysis and Report Writing

Module 24:

ANALYSIS OF QUALITATIVE DATA

STEPS IN DATA ANALYSIS AND REPORT WRITING

Questions you must ask	Steps you will take*	Important elements of each step
What data have been collected for each research objective? Are data complete, accurate?	Prepare data for analysis	Review field experiences Inventory data for each objective/study population Sort data and check quality Check computer outputs
What do the data look like?	Describe variables	Frequency distributions Figures, means
How can the data be summarized for easy analysis?	Cross-tabulate quantitative data	Cross-tabulate in relation to objectives Graphic displays, narratives
	Summarize qualitative data	
For quantitative data: does each research objective aim to describe, compare, or find associations?	Determine the type of statistical analysis required	Review objectives, study type, and variables Statistical description of variables Choosing significance tests
1. How can the data be described?		
2. How can differences between groups be determined?	Analyze paired and unpaired observations	**Student's t-test **Paired t-test **Chi-square test **McNemar's chi-square test
3. How can the associations between variables be determined?	Measure associations between variables	**Scatter diagram **Regression line and correlation coefficient **Relative risk, odds ratio
How should the report be written?	Write the report and formulate recommendations	Prepare outline for report Draft and redraft Summarize findings Summarize conclusions for each objective Formulate recommendations Prepare abstract
How should the findings and recommendations be presented and disseminated?	Present a summary and draft a plan for implementation of recommendations	Discuss summaries with different target groups Discuss plan for implementation

* These steps need not be in the sequence in this diagram. The sequence may be adjusted according to the needs of the workshop.
** These elements are optional and may be omitted if not relevant in a particular course.

Module 24: ANALYSIS OF QUALITATIVE DATA

OBJECTIVES

At the end of this session you should be able to:

1. **Describe** efficient ways of ordering and summarizing qualitative data.
2. **Indicate** why it is essential to start summarizing and analyzing qualitative data during field work.
3. **List** the major steps in analyzing qualitative data and drawing conclusions.
4. **Make an outline** of how you will proceed with the ordering and summarizing of your qualitative data, and with the subsequent analysis.
5. **Indicate**, either now or at the end of data analysis, what additional activities you will undertake to test or confirm your findings to prove their validity.

I. **Introduction**

II. **Procedures for Processing and Displaying Qualitative Data**

III. **Drawing and Verifying Conclusions**

IV. **Strategies for Testing or Confirming Findings to Prove Validity**

I. INTRODUCTION

In previous **Modules (9, 10, 13)** it was pointed out that we use qualitative research techniques if we wish to obtain insight into certain situations or problems concerning which we have little knowledge. Qualitative techniques such as the use of open-ended questions, loosely structured interviews, (focus) group discussions, and observations will, therefore, be appropriate in many studies, especially at the onset.

In health systems research the phase in which recommendations for implementing the research results are drafted provides another good opportunity for applying qualitative techniques.

Qualitative research techniques are often used to describe in depth certain procedures, beliefs, and knowledge related to health issues among the population as well as among health staff. These techniques are also well suited for exploring the reasons for certain behaviour or the opinion of respondents on certain issues. Irrespective of how and for what purpose the data has been collected, the researcher usually ends up with a substantial number of pages of **written text** that need to be analyzed.

Although procedures and outcomes of qualitative data analysis differ from those of quantitative data analysis, the principles are not so different. In both cases the researcher will have to:

- describe the sample population,
- order and reduce or code the data (data processing),
- display summaries of data in such a way that interpretation becomes easy,
- draw conclusions, and
- develop strategies for testing or confirming findings to prove their validity.

We will now examine each of these points in more detail.

II. PROCEDURES FOR PROCESSING AND DISPLAYING QUALITATIVE DATA

Description of the Sample Population

A useful first step in data processing (as well as in reporting findings) is a description of the informants. If numbers allow, relevant background data may be tabulated, for example on age, sex, occupation, education, or marital status, as is the practice in quantitative studies.

However, as qualitative data originate from small samples (sometimes a handful of key informants or focus group discussions) more information is required to place the data in their context.

> **For example**, who were the key informants, what made them qualify as such? Who took part in the focus group discussions? How representative were the participants of the groups they represented? Under what circumstances were observations carried out? Who was observed (and who was not)? What were the reactions of those observed?

Unless this type of information is provided, interpretation of data may be haphazard.

Ordering and Coding Data

Previous modules (**Modules 10.B and 13**) stressed that immediately after each interview or focus group discussion the raw field notes (and tapes, if used) have to be transformed into a well organized set of notes. This version of the notes should reflect as closely as possible what was discussed, but should also include your own observations and comments.

When analyzing such field notes we usually discover that, no matter how good our guidelines for the discussion were, the notes contain valuable data, but also a number of less essential or even useless details. In addition, the data are usually not presented in the order we need them for our analysis, because informants may jump from one topic to the other.

To make the analysis easy, we have to **order** and **reduce** the data. Ordering is best done in relation to the objectives or discussion topics.

If the data are limited and straightforward, we can simply mark all relevant data that belong to the same discussion topic by writing the number of that topic in the margin. If the data are complicated and bulky, it may be helpful to use codes for ordering the data.

Unlike quantitative data, where codes are usually numbers, the codes for qualitative data are usually **labels** which can be easily remembered.

> **For example**, in the focus group discussion on changing weaning practices presented in **Module 10C**, we might code the data in the following way:
>
> | | | |
> |---|---|---|
> | Typ Bab F | - | Types of baby food |
> | Freq Bab F | - | Frequency of baby feeding |
> | Ag Ons SF | - | Age of onset of soft food |
> | Typ SF | - | Types of soft food |
> | Freq Prep SF | - | Frequency of preparation of soft foods |
> | Way Prep SF | - | Way in which the soft foods are prepared |
> | Reas Earl SF | - | Reasons for early introduction of soft foods, etc. |

The codes will closely follow the topics of the discussion guide or of the checklist for observations, although you may have several codes for each topic.

Notes:

It is advisable to do the coding in pencil, as you may want to further refine or change your codes as you proceed with the analysis. Put the codes in the left margin.

You can use the right margin of your field note report for remarks that come up when reading through your field notes; conclusions, incompleteness of data, further questions, or even topics to be added.

Summarizing Data: Graphic Displays in Charts and Figures

After ordering the data we will have to summarize them. The first step is listing the data that belong together. If we have coded the data, we list all the data that have been given the same code.

We may try to list the data in a still more condensed form and in such a way that it becomes easy to answer our research questions.

> **Note:**
>
> When listing data remember to identify the source for each item (interview number or field note page number, for example), so that you can always go back to the information in its original context.

It is possible to further summarize the data **GRAPHICALLY** in a chart (e.g., a matrix) or a figure (e.g., diagram, flow chart). This will be particularly helpful when there is a need to interpret large amounts of data.

Matrices

> A MATRIX is a chart that looks like a table, but contains words instead of numbers.

In the focus group discussion on changing weaning practices, the researchers listed the answers of young mothers concerning the introduction of soft foods and those of mothers above child-bearing age using the codes as outlined above. They then summarized these answers in a matrix:

Figure 24.1. Matrix on introduction of soft baby foods among mothers of different age groups.*

AGE GROUPS	ONSET SOFT FOOD	TYPE OF FOOD	FREQUENCY OF SF/DAY
Young mothers (20-30)	range: 4-7 months average: 6 months	• soft porridge • soft porridge with pound groundnuts • mashed potatoes, mashed fruits, soaked biscuits	1-2 times daily • depends on availability of mother and caretaker • depends on appetite of child
Mothers past child-bearing age (>45)	range: 5-11 months average: 8.5 months	• soft porridge • soft fruits	1-2 times daily • depends on availability of mother and caretaker • depends on appetite of child

* Youngest child of young mothers; oldest child of mothers above child bearing age.

This type of display made it easy for the researchers to conclude that:

- younger mothers start giving soft foods, on average, 2.5 months earlier than the generation of their own mothers;
- younger mothers use a larger variety of soft weaning foods than older mothers; and
- the two generations do not seem to differ in the frequency with which they give/gave soft foods to their babies.

In other words, by displaying their results in graphic form they were already analyzing them.

Matrices are the most common form of graphic display of qualitative data. They can be used to order information in many ways; for example, according to:

- time sequence (of procedures being investigated, for example),
- type of informants (as in the example above),
- location of data collection,
- type of activity, or
- reasons for certain behaviour.

The researchers can also include their own comments in a matrix (for example on quality of observed health staff performance).

Diagrams

> A DIAGRAM is a figure with boxes or circles containing variables and arrows indicating the relationships between the variables.

The focus group discussion on changing weaning practices might provide information on reasons for early or late weaning that we could satisfactorily summarize in diagrams, for example as shown in the two diagrams below.

Figure 24.2. Reasons for early introduction of soft foods by young mothers.

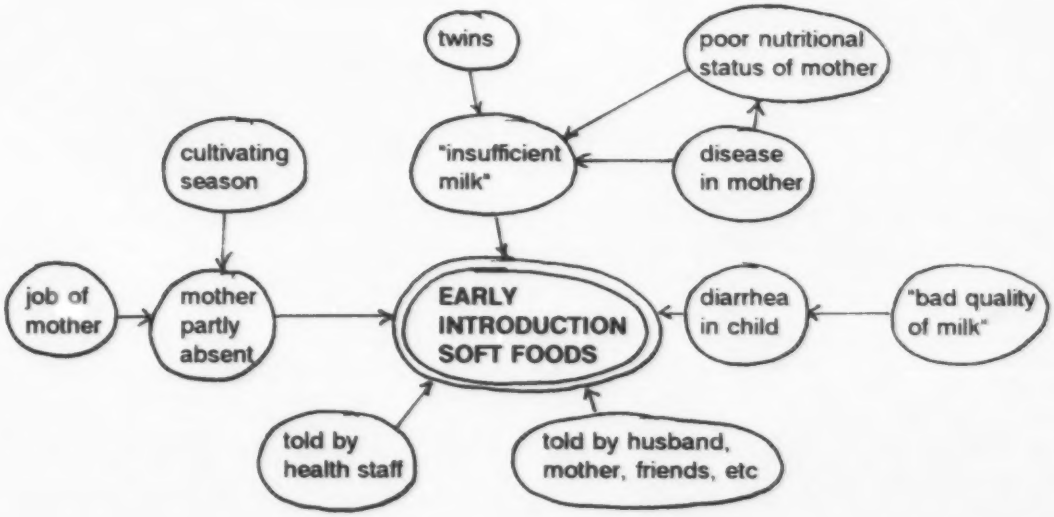
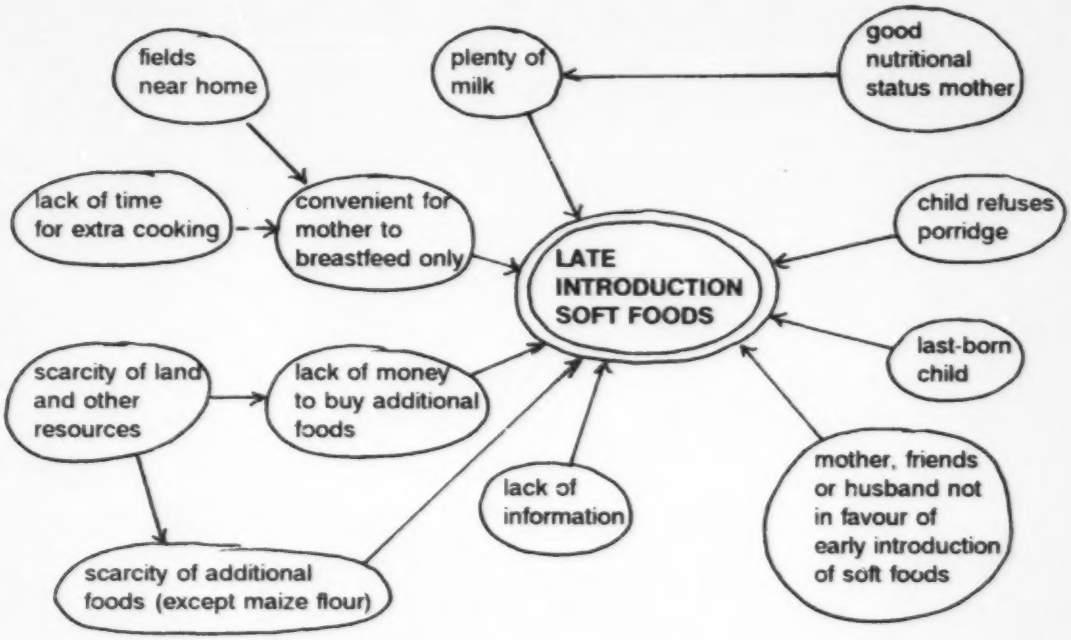


Figure 24.3. Reasons for late introduction of soft foods by young mothers.


fields near home
plenty of milk
good nutritional status mother
lack of time for extra cooking
convenient for mother to breastfeed only
child refuses porridge
LATE INTRODUCTION SOFT FOODS
last-born child
scarcity of land and other resources
lack of money to buy additional foods
lack of information
mother, friends or husband not in favour of early introduction of soft foods
scarcity of additional foods (except maize flour)

Such diagrams can be prepared after one focus group discussion, for example, for young mothers as well as for mothers above child-bearing age. It will help to explore the problem even further in subsequent discussions. **For example**: what influences whether mothers, friends or husbands are in favour or not in favour of the early introduction of soft weaning foods? How it is possible that some mothers have information on what soft foods to provide and how to prepare them, whereas other mothers lack this information? How could this information best be provided?

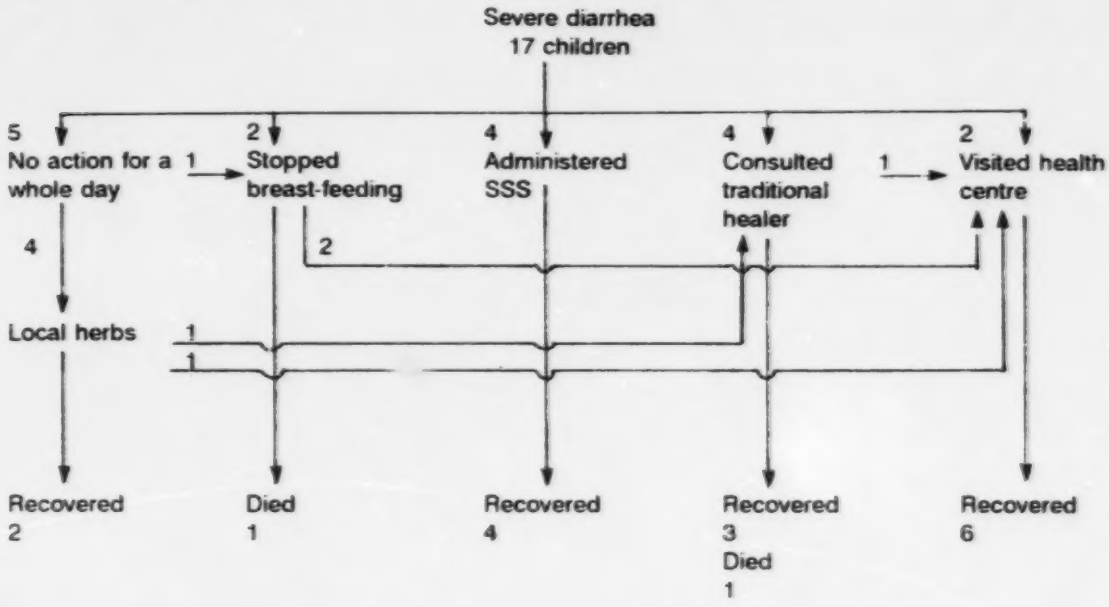
Some diagrams may be more complex than the ones presented here, especially if you include the attitudes and opinions of informants. You may change and elaborate the diagrams as you receive more data or proceed with the analysis of existing data. Diagrams, like matrices, are of great assistance in keeping an overview of the data collected and in guiding data analysis.

Flow Charts

> FLOW CHARTS are special types of diagrams that express the logical sequence of actions or decisions.

The figure in **Module 21** indicating the successive steps in data analysis is an example of a flow chart. Another example, summarizing actions taken by mothers whose children have diarrhea is presented below.

Figure 24.4. Action taken by 17 mothers whose breastfed children (6-18 months old) had severe diarrhea.



From the flow chart we can see that different actions were taken by the 17 mothers who were interviewed. It is remarkable that those children who were given salt-sugar-solution (four in total) all recovered without any other form of treatment. Three mothers stopped breastfeeding their children, and one of these children died. The other two recovered after being taken to a health centre, etc, etc.

Tables

A TABLE is a chart with rows and columns containing data in the various cells or boxes.

Sometimes qualitative data can be categorized, counted and displayed in tables. Answers to open-ended questions in questionnaires can often be categorized and summarized in this way. You may want to analyze the content of the individual answers as well (see Module 21).

Narrative Text

An important part of the presentation of qualitative data will consist of narrative text. The reason why we first discussed graphic display of the findings in charts and figures is that researchers are tempted to go in all directions and lose themselves in details if they start writing straight away. Graphic displays such as matrices, diagrams, flow charts and tables will help you stay on track and be as succinct as possible. You may present them either in the text or in appendices.

III. DRAWING AND VERIFYING CONCLUSIONS

Drawing and verifying conclusions is the essence of data analysis. It is not an isolated activity, however. When we start summarizing our data, we continuously draw conclusions, and modify or reject quite a number of them as we proceed. Writing helps generate new ideas as well. Therefore **writing should start as early as possible**, right from the onset of data processing and analysis.

> **Note:**
>
> Collection, processing, analysis and reporting of qualitative data are closely intertwined, and not (as is the case with quantitative data) distinct successive steps. For that reason it should always be possible (and often is essential) to go back to the original field notes and to the field to verify conclusions, collect additional data, and get feedback from all parties concerned.

Keeping this in mind, let us briefly recapitulate the various possible steps in data processing and analysis that lead to the drawing of justified conclusions:

1. **Ordering** the data in relation to the objectives/research questions.

2. **Categorizing** or labeling answers that have similar characteristics or patterns.

3. **Displaying the summarized data** in charts and figures such as matrices and diagrams or tables **to visualize possible relationships between certain variables.**

   Note that developing matrices is a particularly useful way of discovering such relationships.

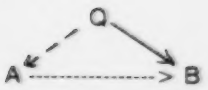
4. **Identifying variables and associations between variables**

   The researcher who follows a qualitative approach should be like a detective who searches for evidence, accounts for countervailing evidence, verifies the findings by looking for independent, supporting evidence, until he is confident about possible associations and causal links between certain variables.

   For example, if we find that among the mothers who wean their children early there are quite a number with jobs, we may assume that having a job contributes to early weaning. Similar studies carried out elsewhere with similar findings support this assumption (independent evidence). Only if there are very few women with jobs among the mothers who wean their children late, however, can we be more certain that our assumption is true, and for each of those exceptions we should try to find an explanation. Do the mothers take their children with them (creche at place of work) or do they work near their homes so that they can feed the baby during breaks? Or do they successfully combine breastmilk with alternatives? If yes, why do not more mothers try this combination? etc. etc.

5. **Finding confounding or intervening variables**

   Sometimes variables appear to be related but the association cannot easily be explained. Other times it seems that variables should logically go together, but you cannot find a relationship. In cases such as these it may be that there is another variable, "Q", influencing the association between the two variables concerned, that has to be identified.



One example is the finding in a malaria control programme that villages where homes were sprayed in the afternoon had a higher incidence of malaria than villages sprayed in the morning. (See example 2 in exercise 4a of **Module 9**.) The program managers could not explain this association until, through participatory observation, it was found that the sprayers used most of their DDT in the mornings so that the load to carry in the afternoon would be lighter.

Another example is the relationship between the quality of drinking water and the incidence of diarrhea. One would expect that the incidence of diarrhea would decrease as the number of water faucets in a village increased. If there is no change over time, there might be an intervening variable. People, for example, may dislike the taste of tap-water so much that they use it for everything, except for drinking.

Note:

Such unexplained associations may appear in any study. The essential characteristic of a qualitative research approach is that it purposively looks for such associations **during** the fieldwork, by intertwining data collection and analysis. In strictly quantitative surveys one usually only discovers the associations and gaps in the data needed to explain them afterwards.

6. **Looking for logical chains of evidence**

If a researchers managers to painstakingly establish associations for all (connected) variables that lead to a certain outcome, he builds a "logical chain of evidence." It is especially rewarding to analyze a successful enterprise by looking at the different conditions that in combination have to be fulfilled to obtain the positive outcome. **Health Systems Research is so concerned with problem solving that we may forget that analysis of successes can teach us much about reasons for failure.**

It may not be possible in the limited time we have for research to build a perfect and completely logical chain of evidence to support an observed outcome. It is even less likely that we can move from our causal networks and logical chains of evidence to more generally applicable THEORIES. Nevertheless, to prove the validity of our findings we have to apply the same strategies for testing and/or confirming them as researchers do who attempt to make that last step.

IV. STRATEGIES FOR TESTING OR CONFIRMING FINDINGS TO PROVE VALIDITY

Researchers who use quantitative research designs reduce their data to numbers and apply statistical tests. This does not necessarily ensure that their research results are valid: something may have gone wrong during sampling or collection of data. The following strategies therefore should be used by all researchers. They are particularly relevant, however, to qualitative research, as most of those strategies are used to establish logical coherence in the data collected.

1. **Check for representativeness of data**

Although in qualitative research informants have usually not been selected randomly, they must have been selected systematically, according to previously established rules. (See **Module 11**.) Check

whether you have indeed interviewed all categories of informants needed to get a complete picture of your topic (not relying excessively on talkative authorities). Make sure that you do not generalize from unrepresentative events.

2. **Check for bias** due to observer bias or the influence of the researcher on the research situation. We discussed this in detail in **Module 10A.**

3. **Cross-check data with evidence from other, independent sources**

These sources may be different independent informants, different research techniques employed to investigate the same topic, or results from other, similar research studies. (See **Modules 5, 9, and 10.**) The data should confirm or at least not contradict each other.

Actively cross-checking data, looking for independent evidence or counter-evidence, is one of the most important ways a researcher can enhance the validity of his data.

> **For example**, answers of husbands and wives (and other informants concerned) should confirm each other on such issues as who decides whether and what family planning methods should be used, who decides whether daughters should be circumcised, or who keeps the household money.

4. **Compare and contrast data**

This can be built into the research design through including different categories of informants.

> If we want to be sure, **for example**, that variable A (high level of education) influences variable B (use of family planning methods) we have to compare a group of mothers with high education to a group of mothers with low education. We expect to find a lower use of family planning methods by the second group.

Comparing and contrasting data is important if you are attempting to **identify** your variables as well as confirm associations between variables.

5. **Use extreme (groups of) informants to the maximum**

In the discussion of study design and sampling we stated that it may be useful to look for categories of informants that represent the extremes on a certain variable.

> **For example**, you may find it most useful to study heavy defaulters and regular attenders of TB services, leaving out the category of irregular attenders. This may be the quickest way of identifying the key variables that influence the behaviour of TB patients.

6. **Do additional research to test the findings of your study**

You may have little time to do additional research in the field work period between the two workshops, but the results of your study may be so intriguing that you decide to do a follow-up study afterwards. Such a study may be undertaken for several reasons:

- to replicate certain findings,
- to rule out (or identify) possible intervening variables,
- to rule out rival explanations by investigating them, or
- to look for negative evidence.

Additional studies undertaken for one or more of these reasons may serve to make the results of your original study more convincing.

7. **Get feedback from your informants**

Throughout **Modules 1-20** we have stressed that you need to involve all parties concerned in the various stages of the research. This is important not only for ethical reasons or because it will improve the chances that the results will be implemented, but also because **it will improve the quality of your proposal, your data, and the conclusions drawn from these data**. Suggestions and additional information collected during feedback sessions will invariably increase the quality of your research report.

GROUP WORK (Time needed will depend on the amount of qualitative data collected.)

- Check whether you listed all sources of qualitative data for each objective when, in the group work session of **Module 21**, you made an inventory of all your data.
- If your study included (groups of) informants from which you collected qualitative data only (e.g. focus group discussions or key informants), describe your sample.
- If you have many pages of qualitative data, code it and then list statements or observations for each code.

  Otherwise, list the data for each objective.
- Decide whether you should use matrices, diagrams and/or narrative text to summarize your data.
- Draw conclusions after checking and cross checking your findings. State whether and how you would like to further test certain conclusions.

References

Miles, M.B., Huberman, A.M. 1984. Qualitative data analysis, a sourcebook of new methods. Sage Publications, Beverly Hills, CA, USA.

Spradly, J.P. 1979. The ethnographic interview. Holt, Rinehart and Winston, New York, NY, USA.

Walker, R. (ed). 1985. Applied qualitative research. Gower Publishing Company Ltd, Hants, UK.

Yin, R.K. 1984. Case study research: design and methods. Sage Publications, Beverly Hills, CA, USA.

NB: The main source of inspiration for writing this module was Miles and Huberman. Sections III and IV of this module are heavily abbreviated and adapted versions of their chapter VII.

Trainer's Notes

Module 24: ANALYSIS OF QUALITATIVE DATA

Timing and Teaching Methods

1 hour Introduction and discussion

1 hour+ Group Work

(Time should be adjusted, depending on amount and type of qualitative data)

Introduction and Discussion

- If none of the groups has qualitative data other than some open-ended questions in questionnaires, you may wish to concentrate on sections I and II, to give participants an overview of how one could process qualitative data, and only briefly touch on sections III and IV.

- However, if the course participants have experience/training/interest in research, you might fully cover sections III and IV, even if they have not collected large quantities of qualitative data. The procedures presented for drawing conclusions and testing validity are pertinent to all types of research, and the methods for checking and cross-checking data may not be known to all participants.

- Refer to the analysis diagram each group made when preparing its research proposal, to the flow chart in front of **Modules 1-18** and **22-32**, and present any other examples of charts or graphs you can provide on overhead sheets or flip charts.

- If one or more groups have done extensive qualitative research, cover the module in detail, using examples from their studies. Most likely none of the participants will be very familiar with analysis of qualitative data.

- Let the groups who have done qualitative research describe in plenary how they analyzed the data from focus group discussions, observations, and/or interviews with key informants. Ask what additional questions they added or questions they dropped in the course of successive interviews, and why.

Group Work

- **For all groups:**

  Check whether any of their open-ended questions require analysis of the content of individual answers. Some opinion questions might provide valuable illustrative material for their reports.

Take note of this, as the groups might forget such data when they get involved in tables and statistics.

Discuss whether merely listing the statements is sufficient for content analysis or whether graphical display of the data would be desirable.

- **For groups that have qualitative data in addition to answers to open-ended questions:**

  Review all the data available with them and ask them to follow the group work directions.

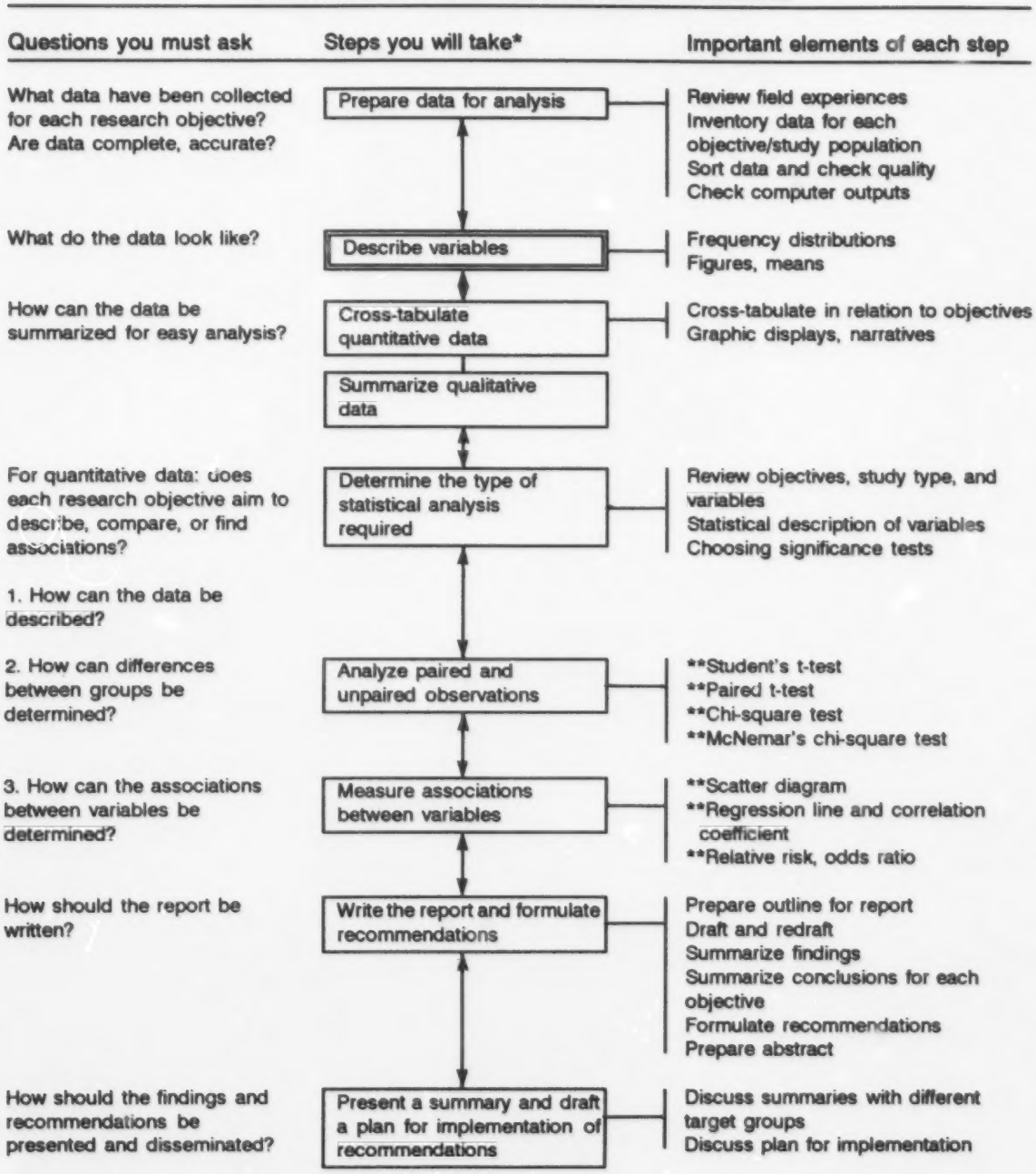
Health Systems Research Training Series

Volume 2, Part II: Data Analysis and Report Writing

Module 25:

DESCRIPTION OF VARIABLES: PART II

STEPS IN DATA ANALYSIS AND REPORT WRITING


Questions you must ask
Steps you will take*
Important elements of each step
What data have been collected for each research objective? Are data complete, accurate?
Prepare data for analysis
Review field experiences
Inventory data for each objective/study population
Sort data and check quality
Check computer outputs
What do the data look like?
Describe variables
Frequency distributions
Figures, means
How can the data be summarized for easy analysis?
Cross-tabulate quantitative data
Cross-tabulate in relation to objectives
Graphic displays, narratives
Summarize qualitative data
For quantitative data: does each research objective aim to describe, compare, or find associations?
Determine the type of statistical analysis required
Review objectives, study type, and variables
Statistical description of variables
Choosing significance tests
1. How can the data be described?
2. How can differences between groups be determined?
Analyze paired and unpaired observations
**Student's t-test
**Paired t-test
**Chi-square test
**McNemar's chi-square test
3. How can the associations between variables be determined?
Measure associations between variables
**Scatter diagram
**Regression line and correlation coefficient
**Relative risk, odds ratio
How should the report be written?
Write the report and formulate recommendations
Prepare outline for report
Draft and redraft
Summarize findings
Summarize conclusions for each objective
Formulate recommendations
Prepare abstract
How should the findings and recommendations be presented and disseminated?
Present a summary and draft a plan for implementation of recommendations
Discuss summaries with different target groups
Discuss plan for implementation

* These steps need not be in the sequence in this diagram. The sequence may be adjusted according to the needs of the workshop.

** These elements are optional and may be omitted if not relevant in a particular course.

Module 25: DESCRIPTION OF VARIABLES, PART II

OBJECTIVES

At the end of this session you should be able to:

1. Explain what is meant by a range, a percentile, a standard deviation, a normal distribution, a standard error, and a 95% confidence interval.
2. Calculate ranges, standard deviations, standard errors, and 95% confidence intervals for your own data, where appropriate.
3. Interpret the results of these calculations.

I. Introduction

II. Measures of Dispersion

III. The Normal Distribution

IV. The Relation of the Sample to the Whole Population

V. How to Determine the Extent to Which the Sample Represents the Population as a Whole

I. INTRODUCTION

As we have seen, the mean, median, and mode are measures of the central tendency of a variable, but they do not provide any information of how much the measurements vary. This module describes some common measures of variation (or variability), which in statistical textbooks are often referred to as measures of dispersion. Also, the concepts of normal distribution, standard error, and confidence interval are introduced. We will need these concepts when we use statistical tests.

II. MEASURES OF DISPERSION

Range

> The **RANGE** of a set of measurements is the difference between the smallest and the largest measurement.

For example, if the weights of 7 pregnant women were 40, 41, 42, 43, 44, 47, and 72 kg, the range would be 72 - 40 = 32 kg.

Although simple to calculate, the range does not tell us anything about the distribution of the values between the two extreme ones.

If the weights of 7 other pregnant women were 40, 46, 46, 46, 50, 60, and 72 kg the range would also be 72 - 40 = 32 kg although the values are very different from those of the previous example.

Percentiles

A second way of describing the variation or dispersion of a set of measurements is to divide the distribution into percentiles. As a matter of fact the concept of percentiles is just an extension of the concept of the median, which may also be called the 50th percentile.

> **PERCENTILES** are points which divide all the measurements into 100 equal parts.

The 3rd percentile (P3) is the value below which 3% of the measurements lie.

The 50th percentile (P50) or the median is the value below which 50% of the measurements lie.

The concept of percentiles is used by nutritionists to develop standard growth charts for specific countries.

Standard Deviation

To determine how much our measurements differ from the mean value there is a measure which we need when we use statistical tests. This measure is called the standard deviation.

The STANDARD DEVIATION is a measure which describes how much individual measurements differ, on the average, from the mean.

To obtain the standard deviation of a set of measurements you have to complete the following steps:

1. Calculate the mean of all the measurements.
2. Calculate the difference between each individual measurement and the mean.
3. Square all these differences.
4. Take the sum of all squared differences.
5. Divide this sum by the number of measurements minus one.
6. Finally (since the differences from the mean have been squared), take the square root of the value obtained (in order to get back to the same unit of measurement).

Example 1:

11 children of 3 years of age were weighed and the following weights were obtained:

13, 14, 14, 15, 16, 16, 16, 17, 17, 18, and 20 kg

The number of measurements (n) is 11.

To calculate the standard deviation:

1. We first calculate the mean: the mean value is 16 kg.
2. Next, we calculate the distance of each measurement from the mean:

   3, 2, 2, 1, 0, 0, 0, 1, 1, 2, 4 (i.e., 16 - 13 = 3; 16 - 14 = 2; etc.)

3. These values are then squared:

   9, 4, 4, 1, 0, 0, 0, 1, 1, 4, 16 (i.e., 3 x 3 = 9; 2 x 2 = 4; etc.)

4. The sum of these squared differences is 40. (i.e., 9 + 4 + ...)
5. This sum is divided by the number of measurements minus one ($n - 1 = 10$):

   $40 / 10 = 4$

6. Finally, we take the square root to obtain the standard deviation from the mean:

   $\sqrt{4} = 2$ kg

A large standard deviation shows that there is a wide scatter of measured values around the mean, while a small standard deviation shows that the individual values are concentrated around the mean with little variation among them.

For instance, if the weights in **Example 1** were

10, 11, 12, 14, 16, 16, 16, 18, 20, 21, and 22 kg

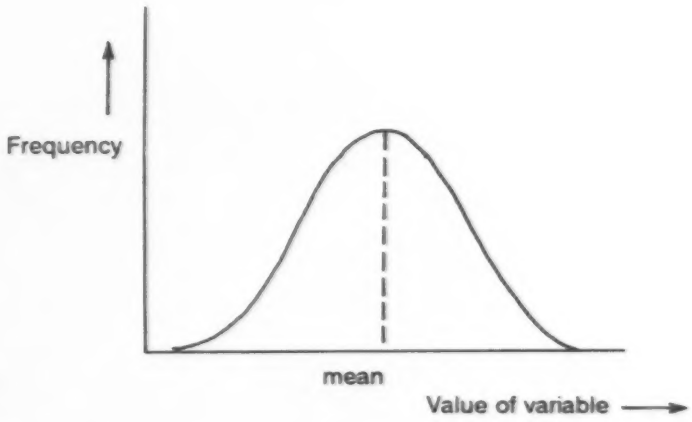
the mean weight would still have been 16 kg. However, the standard deviation would have been 4 kg, indicating a much larger variation in the observations.

Another way to calculate a standard deviation is to use the formula which is given in **Annex 25.1**. Fortunately many pocket calculators can do this calculation for us, but it is still important to understand what it means.

III. THE NORMAL DISTRIBUTION

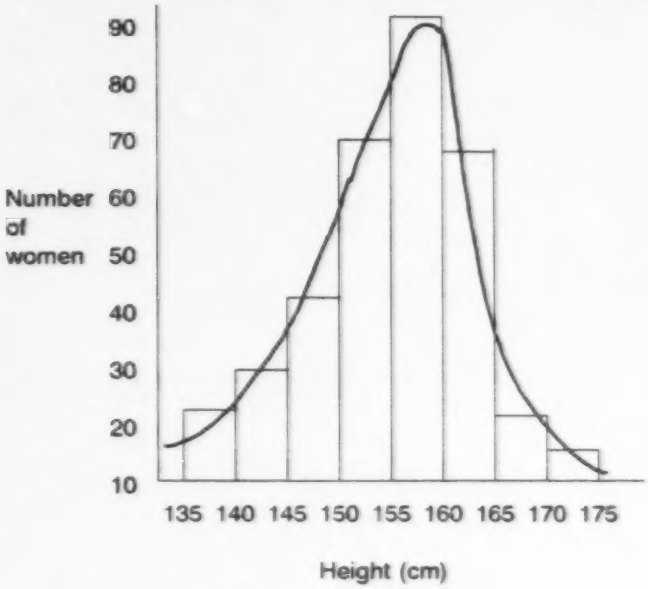
Many variables have a **normal distribution**. This is a bell shaped curve with most of the values clustered near the mean and a few values out near the tails. The normal distribution is symmetrical around the mean. The mean, the median and the mode of a normal distribution have the same value.

Figure 25.1. Normal distribution curve.



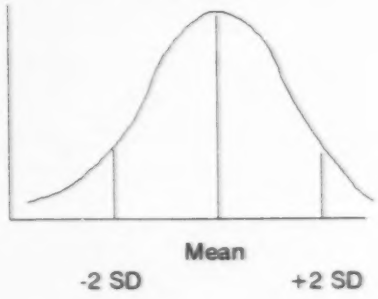
In the following figure a histogram of the heights of pregnant women attending an antenatal clinic is shown, with the normal curve drawn over it.

Figure 25.2. Heights of pregnant women attending an antenatal clinic.



An important characteristic of a normally distributed variable is that 95% of the measurements have values which are approximately within 2 standard deviations (SD) of the mean.[2] This is shown in the figure below.

Figure 25.3. A normally distributed variable.



Example 2:

If the mean height of a group of 120 women is 158 cm and the standard deviation is 3 cm, it means that 95% of the women are between 152 and 164 cm (assuming that the heights are normally distributed). In other words, 2.5% of the women (which in this case corresponds to 3 women) are shorter than 152 cm and 2.5% (or 3 women) are taller than 164 cm.

---

[2] To be more precise, 95% of the measurements have values which are within 1.96 standard deviations of the mean.

Many statistical tests require that the variables be normally distributed. Therefore it is important to examine the frequency distributions for your variables to determine which of them do not have normal distributions. Out of the examples given in the modules so far there are several which are not normally distributed (e.g. **Example 4** in **Module 22**).

IV. THE RELATION OF THE SAMPLE TO THE WHOLE POPULATION

When you undertake a study, it is usually necessary to draw a sample from the study population. You will then describe the population on the basis of the information collected from the sample. In other words, you will try to generalize the findings from the sample to the larger population. Obviously, this can only be done if the sample is selected in such a way that it can be considered representative of the whole population.

Any value of a variable obtained from the sample (e.g. a mean) can then be considered as an estimate of the corresponding population value.

> **For example**, if 158 cm is the calculated mean height of a sample of 120 women you hope it is a good approximation of the mean height of the population of women as a whole.

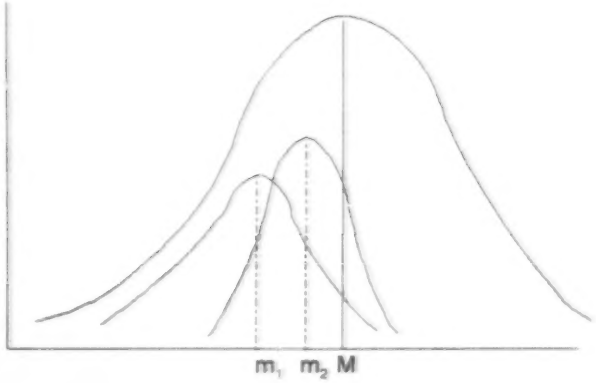
However, the sample mean is not likely to be identical to the population mean.

> If you draw another sample of 120 mothers, you might find a mean of 157 cm, which is not identical to the first sample mean. It probably also differs from the true mean height of the total population from which the sample was drawn.

This phenomenon is called **SAMPLING VARIATION.**

The following figure shows a frequency distribution curve of a population with the curves of two different samples inside it.

Figure 25.4. A frequency distribution curve and two sample curves for a variable that is normally distributed.



M = population mean
m_1 = mean of sample 1
m_2 = mean of sample 2

Note that any representative sample will have a distribution curve similarly shaped to the population curve, but it can fall anywhere within the population curve.

V. HOW TO DETERMINE THE EXTENT TO WHICH THE SAMPLE REPRESENTS THE POPULATION AS A WHOLE

To find out to what extent a particular sample value deviates from the population value, a range or an interval around the sample value can be worked out which will most probably contain the population value.

This range or interval is called the **CONFIDENCE INTERVAL.**

> A CONFIDENCE INTERVAL is the interval or range of values which most likely encompasses the true population value. The lower and upper limits of this interval are termed **confidence limits.**

For example:

A 95% confidence interval of 152 to 164 cm for the mean height of a population of women means that you are 95% certain that the real population mean, which you cannot know exactly unless you measure the heights of all women, lies between 152 and 164 cm. (152 cm is the lower confidence limit, 164 cm is the upper confidence limit.)

The calculation of a confidence interval takes into account the **STANDARD ERROR**. The standard error gives an estimate of the degree to which the sample mean varies from the population means. It is computed on the basis of the standard deviation.

We will now discuss how to calculate:

- the standard error and the 95% confidence interval of a mean (for numerical data), and
- the standard error and the 95% confidence interval of a percentage (for categorical data).

Standard Error and 95% Confidence Interval of a Mean

When dealing with numerical data you may wish to estimate to what degree the sample mean varies from the population mean.

The standard error for the mean is calculated by dividing the standard deviation by the square root of the sample size:

$$\text{standard deviation} / \sqrt{\text{sample size}} \text{ or } SD / \sqrt{n}$$

It can be assumed, for a normally distributed variable, that approximately 95% of all possible sample means lie within two standard errors of the population mean. In other words, we can be 95% sure that the population mean, of which we want to have the best possible estimate, lies within two standard errors of our sample mean.

When describing variables statistically you usually present the calculated sample mean plus or minus two standard errors. This is then called the **95% CONFIDENCE INTERVAL**. It means that you are about 95% certain that the population mean is within this interval.

In **Example 1** the weights of a random sample of 11 three year old children were taken. The sample mean was 16 kg and the standard deviation of the sample was 2 kg.

The standard error is

$$2 / \sqrt{11} = 0.6 \text{ kg}$$

The 95% confidence interval is

$$16 \pm (2 \times 0.6) \quad = \quad 14.8 \text{ to } 17.2 \text{ kg}$$

This means that we are approximately 95% certain that the mean weight of all three year old children in your population lies between 14.8 and 17.2 kg.

Note that the larger the sample size, the smaller the standard error and the narrower the confidence interval will be. Thus the advantage of having a large sample size is that the sample mean will be a better estimate of the population mean.

If in the above example the sample size was 20 (instead of 11), the standard error would have been:

$$2 / \sqrt{20} = 0.45 \text{ kg}$$

and the 95% confidence interval for the mean weight would have been from 15.1 to 16.9 kg.

Standard Error and 95% Confidence Interval of a Percentage

In the previous section we calculated the standard error and the 95% confidence interval of a sample mean, starting with **numerical data**. We will now do the same for a percentage that was calculated from **categorical data**.

Example 3:

Among a sample of 120 TB patients, which was drawn from the total population of TB patients in the country, it was found that 28 (or 23.3%) did not comply with their out-patient treatment. The other 92 (or 76.7%) exhibited a satisfactory degree of compliance. We now want to calculate the standard error of the percentage of non-compliers (23.3%). This is done as follows.

If p represents one of the percentages (23.3%) and 100 - p represents the other (76.7%), then the standard error of the percentage is obtained by multiplying them, dividing the result by the number in the sample and taking the square root.

The formula for the standard error of a percentage is:

$$\sqrt{p(100 - p) / n}$$

In the **example** this is:

$$\sqrt{23.3 \times 76.7} / 120 = \sqrt{14.89} = 3.9$$

We now also want to calculate the confidence interval for the percentage of non-compliers in the whole country.

The 95% confidence interval is

$$23.3\% \pm (2 \times 3.9) \text{ which is } 15.5\% \text{ to } 31.1\%.$$

This means that there is a probability of 95% that in the population of all TB patients in the country from which the sample of 120 was drawn, 15.5% to 31.1% do not comply with their out-patient treatment.

Note that instead of percentages we can use proportions, which can take on any value between 0 and 1. The formula for the standard error would then be:

$$\sqrt{p(1 - p) / n}$$

In the **example** this is:

$$\sqrt{0.233 \times 0.767 / 120} = 0.039$$

GROUP WORK

- Calculate the range, the standard deviation and the 95% confidence interval for your most important sets of numerical data.

  Interpret the results of these calculations.

- Calculate the 95% confidence interval of percentages for your most important sets of categorical data. This same calculation can be made for numerical data if they are summarized in categories.

  Interpret the results of the calculations.

Save the results of your group work. You will need them when you perform statistical tests.

Annex 25.1. Formula for Calculating Standard Deviation

Standard deviation (SD) $= \sqrt{(\Sigma x^2 - (\Sigma x^2 / n))} / (n - 1)$

where x = each value,
x^2 = the square of each value,
Σ = "the sum of," and
n = the number of observations.

If we apply this formula to **Example 1** (weights of 11 three year old children on page 5), you will find that it is not as difficult as it first seems. The values x are:

13, 14, 14, 15, 16, 16, 16, 17, 17, 18, 20 kg and n = 11

If we square each value we get:

169, 196, 196, 225, 256, 256, 256, 289, 289, 324, 400

The sum of all the squares (Σx^2) is 2856

The sum of all the values (Σx) is 176

Therefore $(\Sigma x)^2$ is $176^2 = 30{,}976$

and $(\Sigma x)^2 / n = 30{,}976 / 11 = 2816$

Standard deviation $= \sqrt{(2856 - 2816)} / 10 = \sqrt{4} = 2$ kg

A faster method to calculate the standard deviation is to use the automatic function built into certain pocket calculators.

Trainer's Notes

Module 25: DESCRIPTION OF VARIABLES, PART II

Timing and Teaching Methods	
1 hour	Introduction and discussion
3 hours*	Group work

Introduction and Discussion

- This is the first module in which statistical concepts are introduced which will be new for some, if not all of the participants, depending on their academic level. Therefore it is recommended that you go slowly, be careful with formulae and use simple examples.

- Be sure that the concept of sampling variation (part IV) is clear to everybody since this is crucial for a good understanding of how significance tests work (next module).

- If participants have little experience with statistics you may decide to leave out section 1 and 2 of Part V from your presentation. However, if one of the groups is doing a descriptive study in which no attempt is made to compare groups, they may need to calculate a standard error of a difference (section 1) or the standard error of a percentage (section 2), instead of performing statistical tests. In that case you should leave in this section when you make your presentation.

Group Work

- If some of the calculations are not appropriate for certain groups they should do at least one of each calculations using sample data for the sake of experience.

- Note that there is no need to calculate standard errors of differences or percentages if statistical tests are to be performed to determine differences between groups or to measure associations between variables. (See **Modules 27-30.**)

Health Systems Research Training Series

Volume 2, Part II: Data Analysis and Report Writing

Module 26:

CHOOSING A SIGNIFICANCE TEST

STEPS IN DATA ANALYSIS AND REPORT WRITING

Questions you must ask	Steps you will take*	Important elements of each step
What data have been collected for each research objective? Are data complete, accurate?	Prepare data for analysis	Review field experiences Inventory data for each objective/study population Sort data and check quality Check computer outputs
What do the data look like?	Describe variables	Frequency distributions Figures, means
How can the data be summarized for easy analysis?	Cross-tabulate quantitative data	Cross-tabulate in relation to objectives Graphic displays, narratives
	Summarize qualitative data	
For quantitative data: does each research objective aim to describe, compare, or find associations?	Determine the type of statistical analysis required	Review objectives, study type, and variables Statistical description of variables Choosing significance tests
1. How can the data be described?		
2. How can differences between groups be determined?	Analyze paired and unpaired observations	**Student's t-test **Paired t-test **Chi-square test **McNemar's chi-square test
3. How can the associations between variables be determined?	Measure associations between variables	**Scatter diagram **Regression line and correlation coefficient **Relative risk, odds ratio
How should the report be written?	Write the report and formulate recommendations	Prepare outline for report Draft and redraft Summarize findings Summarize conclusions for each objective Formulate recommendations Prepare abstract
How should the findings and recommendations be presented and disseminated?	Present a summary and draft a plan for implementation of recommendations	Discuss summaries with different target groups Discuss plan for implementation

* These steps need not be in the sequence in this diagram. The sequence may be adjusted according to the needs of the workshop.
** These elements are optional and may be omitted if not relevant in a particular course.

Module 26: CHOOSING A SIGNIFICANCE TEST

OBJECTIVES

At the end of this session you should be able to:

1. Explain what a significance test is and what its purpose is.
2. Use the tables to choose appropriate significance tests for different sets of data.
3. Choose appropriate significance tests for your own data.

I. **Introduction**

II. **Significance Tests**

III. **How Significance Tests Work**

IV. **Choosing a Significance Test**

I. INTRODUCTION

As said before, the analysis and interpretation of the results of our study must be related to the study objectives. **Module 24** explained how to construct cross-tabulations appropriate to the research objectives.

Suppose we find in a study on smoking behaviour that 30% of the men included in the sample are smokers whereas only 20% of the women are qualified as smokers. How should we interpret this result?

- The observed difference of 10% might be a **TRUE DIFFERENCE**, which also exists in the total population from which the sample was drawn.
- The difference might also be **DUE TO CHANCE**; in reality there is no difference between men and women, but the sample of men just happened to differ from the sample of women. One can also say that the observed difference is due to sampling variation.
- A third possibility is that the observed difference of 10% is due to defects in the study design (also referred to as **BIAS**): with an appropriate study design no such difference would have been found.

In our study design we have attempted to avoid bias, by choosing an appropriate study type, the right data collection techniques, a good sampling procedure, etc. However, it is always worthwhile to consider what biases may have appeared during the implementation of the study that may have influenced the results.

If we feel confident that an observed difference between two groups cannot be explained by bias, we would like to find out whether this difference can be considered as a true difference. We can only conclude that this is the case if we can rule out chance (sampling variation; see **Module 25**) as an explanation. We accomplish this by applying a significance test.

II. SIGNIFICANCE TESTS

> A SIGNIFICANCE TEST estimates the likelihood that an observed study result (e.g., a difference between two groups) is due to chance.
>
> In other words, a significance test is used to find out whether a study result which is observed in a sample can be considered as a result which exists in the population from which the sample was drawn.

Different significance tests have been developed for different sets of data. In this module two flow-charts will be presented to guide you in choosing the appropriate significance tests for your data. The first flow-chart is to be used if you compare groups to detect differences. The second is to be used if you want to measure associations between variables.

In the modules that follow some of the most common significance tests will be discussed in detail. But first let us examine how significance tests work.

III. HOW SIGNIFICANCE TESTS WORK

Behind the use of significance tests is the following reasoning:

- Suppose you observed a difference between two groups (or an association between two variables) in your sample.
- You want to know whether this observed difference between the two groups (or the observed association between the variables) represents a real difference (or real association) in the total study population from which the sample was drawn, or whether it just occurred by chance (due to sampling variation).
- To find this out, you determine how likely it is that your result could have occurred by chance, if in the total population no difference exists between the two groups.

If you are measuring an association between two variables, you determine how likely it is that your result could have occurred by chance (i.e., it occurred due to sampling variation).

If it is **unlikely** (a likelihood of less than 5% or some other predetermined percentage[3]) that your result occurred by chance, you reject the chance explanation and accept that there is a real difference (or association). You may then say that the difference (association) is **statistically significant**.

If it is **likely** (a likelihood of 5% or more) that your result occurred by chance, you cannot conclude that a real difference (association) exists. You then say that the difference (association) is **not statistically significant**.

> **Note:**
>
> In statistical terms the assumption that in the total population **no** real difference exists between groups (or that no real association exists between variables) is called the **NULL HYPOTHESIS**.

Examples of null hypotheses are:

- There is no difference in the incidence of measles between vaccinated and non-vaccinated children.
- Males do not smoke more than females.
- There is no association between families' income and malnutrition in children.

It is important to note that "statistically significant" does not mean that a difference or an association is an important one. The tiniest and most irrelevant difference will turn out to be statistically significant if a big enough sample is taken. On the other hand, a large and important difference may fail to reach statistical significance if a small sample is used.

---

[3] You can choose the p-value yourself: for example, 0.10, 0.05, 0.01, or even 0.001, depending on how sure you want to be that your conclusion is the right one. The choice of 5% is arbitrary. The researcher may decide, for example, on 1% if he or she desires to reduce the likelihood of occurrence by chance.

The likelihood or **PROBABILITY** of observing a result by chance is usually expressed as a **P-VALUE**.

The p-value is expressed as a proportion. A probability of 5% corresponds to a p-value of 0.05.

Therefore, a difference or an association is considered significant if $p < 0.05$. In other words, if the null hypothesis, stating that there is no difference between two groups, is true, you would observe a difference in your data only 5 times or less in every 100 samples examined.

IV. CHOOSING A SIGNIFICANCE TEST

Depending on the aim of your study and the type of data collected, you have to choose an appropriate significance test. This procedure is explained using Tables 26.1 and 26.2.

Note:

Before applying any statistical test, state the null hypothesis in relation to the data to which the test is being applied. This will enable you to interpret the results of the test.

If you are determining differences between groups, use Table 26.1. You need to identify whether you have paired or unpaired observations. In many studies the samples of various groups are selected independently, and no attempt is made to "match" or "pair" the study units in the different samples. In these cases we speak of unpaired samples and unpaired observations.

Sometimes, however, two sets of observations are taken in such a way that the individual observations in one set are paired with the individual observations in the other set. Here we speak of paired samples and paired observations. (See **Modules 9 and 23**.)

Examples of paired and unpaired samples will be given in the following sections, which explain the use of the two tables.

Determining Differences Between Groups (using Table 26.1)

When deciding what test to use to determine whether differences between groups are statistically significant, there are several issues you must consider: first you need to decide whether you have paired or unpaired observations. (See **Modules 9 and 23**, if necessary.) Within each of these categories it is necessary to determine whether the data are nominal, ordinal, or numerical. (For definitions see **Module 22**.)

For **NOMINAL data** (paired or unpaired) the significance test to be used depends on whether the sample is small or large. There is no clear guide to what should be considered "small" or "large." However, in the case of unpaired observations it is better to use **Fisher's exact test** rather than the **Chi-square test** if the total sample is less than 40 *or* if any cell of the table, which must be constructed, has an expected number of less than 5. In these training modules we will only consider the tests used with large samples. The **Chi-square test** will be dealt with in **Module 27** and **McNemar's chi-square test** in **Module 28**.

Example of an unpaired sample:

In a study of the effectiveness of measles vaccination the research team decided to study 100 measles patients aged 1-5 years coming to a clinic and 100 patients in the same age range who did not have measles. When comparing the two groups for their vaccination status, they found that the vaccination rate among measles patients was lower than among non-measles patients. The **Chi-square test** was used to determine whether this difference was statistically significant.

Table 26.1. Choosing a significance test when determining differences between groups.

	Unpaired observations	Paired observations
Nominal data		
Small sample	Fisher's exact test	Sign test
Large sample	Chi-square test* (Module 27)	NcNemar's chi-square test* (Module 28)
Ordinal data		
Two groups	Wilcoxon two-sample test or Mann-Whitney U-test	Wilcoxon signed-rank test
More than two groups	Kruskal-Wallis 1-way analysis of variance	Friedman 2-way analysis of variance
Numerical data		
Two groups	t-test* (Module 27)	Paired t-test* (Module 28)
More than two groups	F-test	

* Tests indicated by asterisks will be discussed in the modules indicated.
Source: Adapted from Riegelman, R.K. 1981. Studying a study and testing a test. Little, Brown and Company, Boston, MA, USA. p. 246.

Example of a paired sample:

In another study with the same objective the researchers thought that age and sex might be important factors affecting the susceptibility to measles and the likelihood of being vaccinated. Therefore, for every measles patient coming to a clinic a non-measles patient of the same age and the same sex was selected from the out-patient queue. Thus, 100 pairs were checked for vaccination status. **McNemar's chi-square test** was used to analyze the data.

For ORDINAL data the significance test to be used depends on whether only two groups or more than two groups are being compared.

The tests to be used for comparing two groups are based on ranking of data: **Wilcoxon's two-sample test**, which gives equivalent results to the **Mann-Whitney U-test**, for unpaired observations and **Wilcoxon's signed-rank test** for paired observations. We will not deal with these tests in our training modules, but if you want to use them, please refer to a textbook on statistics. (See **Annex 26.1**: Chapter 10 of Swinscow's *Statistics at square one* is very clear and easy to understand.) You can also consult your facilitator or a statistician.

Example of a study in which data are ranked:

In a quasi-experimental study to investigate the effect of a health education campaign on the knowledge of management of diarrhea in the community, two groups of villages were selected. The

first group was composed of villages in which the campaign was held, the second of villages in which no health education was given. During analysis the villages were ranked from the highest level of knowledge of adequate treatment of diarrhea to the lowest. **Wilcoxon's two-sample test** was performed to determine whether there was a significant difference between the two groups of villages.

For **NUMERICAL data**, as for ordinal data, the choice of an appropriate significance test depends on whether you are comparing two groups or more than two groups.

In **Modules 27 and 28** we will discuss how to conduct and interpret a **t-test** and a **paired t-test** if you are comparing only two groups. If you are comparing more than two groups, you should refer to a textbook on statistics. (See **Annex 26.1**.)

Example of an unpaired sample, two groups:

In a nutrition study the weights of 142 five year olds living in rural areas and of 171 five year olds living in urban areas were measured. The mean weight for each of the two samples was calculated and compared, using the **t-test**, to determine whether there was a difference.

Example of an unpaired sample, more than two groups:

The mean weights of the following four groups of five year olds were compared: boys living in rural areas, boys living in urban areas, girls living in rural areas, and girls living in urban areas. In this case, the **F-test** was the appropriate choice.

Example of a paired sample, two groups:

The mean weights of adult males and adult females were compared while controlling for height. This meant that for each male of a certain height a female of the same height was selected so that each pair could be compared on weight. The **paired t-test** was used in this instance.

Measuring Associations Between Variables (using Table 26.2)

Table 26.2. Choosing a significance test when measuring associations between variables.

Nominal data	Chi-square test (if sample is large enough)* (Modules 27, 28)	Calculate odds ratio or estimate relative risk* (Module 30)
Ordinal data or numerical data when no linear relationship is suspected	Calculate Spearman's rho or Kendall's tau	Significance of Spearman's rho or Kendall's tau
Numerical data when a linear relationship is suspected	Calculate Pearson's correlation coefficient (r)* (Module 29)	Significance of Pearson's correlation coefficient (r)* (Module 29)

* Tests indicated with an asterix will be discussed in the modules indicated.
Source: Adapted from Reigelman, R.K. 1981. Studying a study and testing a test. Little, Brown and Company, Boston, MA, USA. p. 246.

Determine whether your data are nominal, ordinal, or numerical. If they are numerical, decide whether a linear relationship is suspected. The term "linear relationship" for numerical data means that the association is such that the dependent variable changes in a constant relationship to the independent variable in such a way that the points on a scatter diagram, when joined, are approximated best by a straight line.

For **NOMINAL data** the **relative risk** is a useful measure of association that is often applied in case-control and cohort studies. **Module 30** will deal with calculating the **odds ratio** as an estimate of relative risk in case-control studies (paired and unpaired observations).

Example:

In a case-control study on tuberculosis (TB) in which you looked at type of employment you found that mineworkers were more likely to contract TB than farmworkers. A Chi-square test confirmed that the difference in the incidence of TB between mineworkers and farmworkers was statistically significant. Calculation of the **odds ratio** would help you to express how much more likely mineworkers were to contract TB then farmworkers.

For **ORDINAL data**, **Spearman's rank correlation coefficient (rho)** or **Kendall's tau** can be calculated and tested for significance. If you want to use them, refer to statistics textbook (See Annex 26.1: e.g. Swinscow, 1983, Chapter 12.)

For **NUMERICAL data when a linear relationship is suspected** Pearson's correlation coefficient can be calculated and tested for significance. **Module 29** illustrates how to do this.

Example:

You may want to examine whether the weights of five year olds are associated with their families' income. You suspect that there is a linear relationship between the two variables "family income" and "weight,' such that weight increases with increasing family income.

Note:

An association that is statistically significant does not necessarily imply the existence of a causal relation. However, it often invites further investigation to find out whether a causal relationship does exist.

EXERCISES on Choosing a Test

Using the Tables, identify the appropriate tests for the following research studies.

Exercise 1

A study will be undertaken to compare the effect of a new antihypertensive drug on the diastolic blood pressure of a study group sample compared to the effect of a placebo on an unmatched control group sample (Riegelman 1981, p. 243).

Exercise 2

A study will be conducted to find out whether pregnant women living in households where there is no water supply of their own are at significantly greater risk of experiencing perinatal deaths, and, if so, to measure how strong this association is.

Exercise 3

A study was undertaken to determine whether there was a significant weight loss after a one year course of therapy for diabetes, and whether the amount of weight loss was related to initial weight. The following table gives the initial weights (x) and weights after one year of therapy (y) for 16 newly diagnosed adult diabetic patients.

Initial weight (x) in lb.	Weight after 1 year (y) in lb.	Initial weight (x) in lb.	Weight after 1 year (y) in lb.
140	115	120	123
160	130	145	143
180	135	150	125
120	125	160	140
132	112	160	135
146	130	149	120
190	160	129	119
200	160	150	13

Exercise 4

From previous studies it is determined that 30% of the eligible couples in a Health District practice family planning. After a mass educational program, results indicated that out of 90 eligible couples randomly selected, 40 practiced family planning. The Health Education Officer wishes to know whether his program has had an impact on the target group.

GROUP WORK

Referring to the specific objectives of your study and the list of variables and using the cross-tabulations already made, identify the significance tests you will need to perform on your data.

References

Barker, D.J.P. 1982. Practical epidemiology (3rd ed.). Churchill Livingstone, Edinburgh, UK.

Bradford Hill, A. 1984. A short textbook of medical statistics (11th ed.). Hodder and Stoughton, London, UK.

Fisher, A., Laing, J., Stoeckel, J. 1983. Handbook for family planning operations research design. Population Council, New York, NY, USA.

Kelsey, J.L., Thompson, W.D., Evans, A.S. 1986. Methods in observational epidemiology. Oxford University Press, Oxford, UK.

Kidder, L.H., Judd, C.M. 1986. Research methods in social relations. CBS College Publishing, New York, NY, USA.

Kleinbaum, D.G., Kupper, L.L., Morgenstern, H. 1982. Epidemiologic research - principles and quantitative methods. Van Nostrand Reinhold, New York, NY, USA.

McMahon, B., Pugh, T.F. 1970. Epidemiology - principles and methods. Little, Brown and Co., Boston, MA, USA.

Riegelman, R.F. 1981. Studying a study and testing a test. Little, Brown and Co., Boston, MA, USA.

Schlesselman, J.J. 1982. Case control studies - design, conduct, analysis. Oxford University Press, Oxford, UK.

Swinscow, T.D.V. 1980. Statistics at square one (7th ed.). British Medical Association, London, UK.

Trainer's Notes

Module 26: CHOOSING A SIGNIFICANCE TEST

Timing and Teaching Methods

1/2 hour* Introduction and discussion

1 hour* Group work

* 1/2 to 1 hour should be added to either the plenary or the group work if the trainer decides to ask the participants to complete the exercises on choosing a significance test in one of these two components of the session.

Introduction and Discussion

- This module should not necessarily be presented in full. As stated in the objectives, the main aim of the session is that participants understand **what** significance tests are and **why** they would use them. Also they should be able to USE the tables to choose the appropriate significance tests for different sets of data.

  The examples given for the use of the two tables should **not** be presented.

- The question of **why** you perform significance tests can be introduced by presenting a cross-table (for example the numbers of smokers and non-smokers among males and females) and asking how the difference between males and females (30% versus 20%) can be interpreted. It would even be better to take a cross-table of one of the research teams themselves and ask the same question.

- Once it is clear to everybody why significance tests are performed, you might ask participants to give examples of results (cross-tables) from their own projects for which significance tests have to be performed.

- Stress that there are two tests which are most commonly used: the t-test and the Chi-square test. All other tests mentioned in Table 26.1 are less likely to be used in the types of projects the participants most often develop.

- If your group of participants is advanced enough to learn to use the two tables on their own, you could ask them to complete several or all of the five exercises. They could be asked to take a few minutes during plenary to use the tables to select the appropriate tests, and then volunteers could be asked for the answers. Alternatively, several or all of the exercises could be the first task during group work, with the facilitator playing an active role in assisting group members in becoming adept at using the tables.

- After having introduced this module we recommend that you present part of the next module: either the t-test or the Chi-square test. This will make the theoretical concepts discussed in this module more concrete.

Group Work

- Let the participants decide for which cross-tabulations they should perform significance tests. They should also determine which test is appropriate for each cross-tabulation selected.

Suggested Tests for Research Studies Given in Exercises

Exercise 1:

The study deals with **samples**. We are interested in significant differences between the outcomes or dependent variable, which is change in diastolic blood pressure. This is **numerical data** (Table 26.1).

There is one comparison (two groups) and the samples are **unmatched**. Hence the **t-test** is the appropriate test.

Exercise 2:

We assume that these are **samples**. It is **differences** that are being studied for significance, and the outcome or dependent variable studied is number of perinatal deaths, i.e., **nominal** data (Table 26.1). Such studies are done on large samples and these are **unmatched**. Hence the test is **Chi-square**.

If we want to find the strength of **association** we should use Table 26.2. As this is **nominal data** the **odds ratio** or **relative risk** must be calculated.

The same 2 x 2 table constructed for Chi-square test can be used here also.

Exercise 3:

Samples - differences - dependent variable weights - numerical data (Table 26.1). There was only one sample of 16 patients, but measurements were done twice on each patient. This is thus a **matched (or paired) sample**. In this case it was self-pairing.

Hence the test is **matched t-test**. In regards to whether the weight loss was related to initial weight, it is a test of degree of **association**.

Because the data are **numerical data**, calculate **Pearson's correlation coefficient (r)**. If it is necessary to test statistical significance of association, use Table 26.2 which leads to statistical significance of **Pearson's r**.

Exercise 4:

In this study one **sample** has been selected. It is compared with known population data. Because the data are **nominal**, the test deals with differences between proportions or percentages.

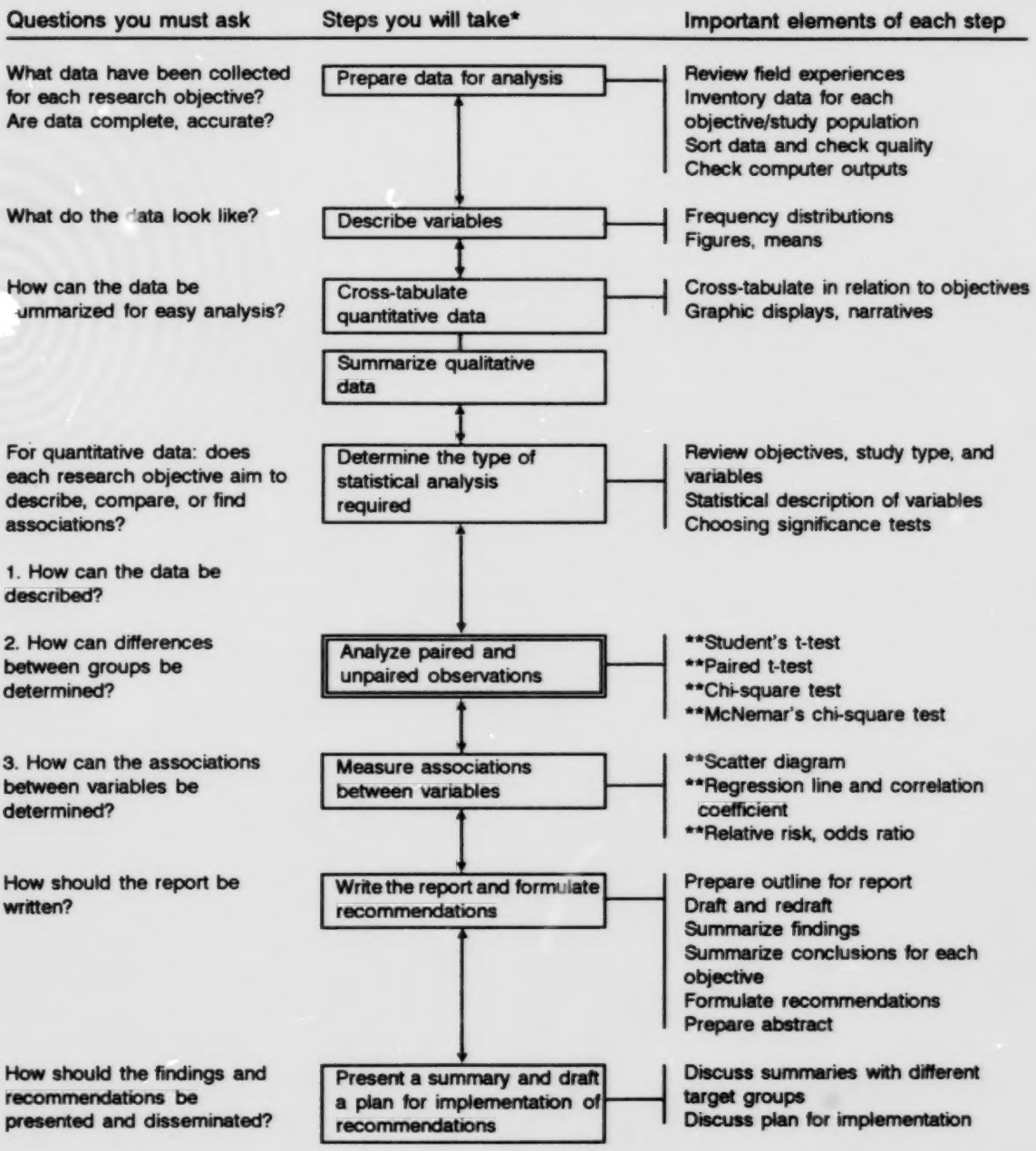
Health Systems Research Training Series

Volume 2, Part II: Data Analysis and Report Writing

Module 27:

DETERMINING DIFFERENCES BETWEEN GROUPS: PART I — ANALYSIS OF UNPAIRED OBSERVATIONS

STEPS IN DATA ANALYSIS AND REPORT WRITING


Questions you must ask
Steps you will take*
Important elements of each step
What data have been collected for each research objective? Are data complete, accurate?
Prepare data for analysis
Review field experiences
Inventory data for each objective/study population
Sort data and check quality
Check computer outputs
What do the data look like?
Describe variables
Frequency distributions
Figures, means
How can the data be summarized for easy analysis?
Cross-tabulate quantitative data
Cross-tabulate in relation to objectives
Graphic displays, narratives
Summarize qualitative data
For quantitative data: does each research objective aim to describe, compare, or find associations?
Determine the type of statistical analysis required
Review objectives, study type, and variables
Statistical description of variables
Choosing significance tests
1. How can the data be described?
2. How can differences between groups be determined?
Analyze paired and unpaired observations
**Student's t-test
**Paired t-test
**Chi-square test
**McNemar's chi-square test
3. How can the associations between variables be determined?
Measure associations between variables
**Scatter diagram
**Regression line and correlation coefficient
**Relative risk, odds ratio
How should the report be written?
Write the report and formulate recommendations
Prepare outline for report
Draft and redraft
Summarize findings
Summarize conclusions for each objective
Formulate recommendations
Prepare abstract
How should the findings and recommendations be presented and disseminated?
Present a summary and draft a plan for implementation of recommendations
Discuss summaries with different target groups
Discuss plan for implementation

* These steps need not be in the sequence in this diagram. The sequence may be adjusted according to the needs of the workshop.
** These elements are optional and may be omitted if not relevant in a particular course.

Module 27: DETERMINING DIFFERENCES BETWEEN GROUPS: PART I — ANALYSIS OF UNPAIRED OBSERVATIONS

OBJECTIVES:

At the end of this session you should be able to:

1. Decide when to apply the unpaired t-test and chi-square test.
2. Calculate t-values and chi-square values.
3. Use the t-tables and chi-squared tables to assess whether the t- and chi-square values are significant.
4. Make a decision concerning whether you can use these tests on your data and, if so, which test should be used on which data.
5. Perform these tests on your data.

I. **Introduction**

II. **T-test**

III. **Chi-square (χ^2) test**

I. INTRODUCTION

When examining the cross-tabulations of your major variables (see **Module 23**) you probably have observed differences between groups. You may want to find out if these differences are likely to be due to chance, or if they are real (statistically significant) differences.

To determine this, you can perform two types of tests. These are:

- t-test, and
- chi-square test.

The **t-test** is used for NUMERICAL data and is used when comparing the means of two groups.

The **chi-square test** is used for CATEGORICAL data and is used when comparing proportions of events occurring in two or more groups.

Both tests are used for UNPAIRED observations. For observations which are PAIRED two different tests are used, again depending on whether the data are categorical or numerical. Information on these tests is given in **Module 28**.

II. T-TEST

The **t-test**, also referred to as **Student's t-test**, is used for numerical data to determine whether an observed difference between the means of two groups can be considered statistically significant.

Example 1:

It has been observed that in a certain province the proportion of women who are delivered through Caesarean section is very high. A study is, therefore, conducted to discover why this is the case. As small height is known to be one of the risk factors related to difficult deliveries, the researcher may want to find out if there is a difference between the mean height of women in this province who had normal deliveries and of those who had Caesarean sections. The null hypothesis would be that there is no difference between the mean heights of the two groups of women. Suppose the following results were found:

Table 27.1. Mean heights of women with normal deliveries and of women with Caesarean sections.

	Number of women included in study	Mean height in cm	Standard deviation
Normal delivery	60	156	3.1
Caesarean section	52	154	2.8

A t-test would be the appropriate way to determine whether the observed difference of 2 cm can be considered statistically significant. To actually perform a t-test you have to complete 3 steps:

1. calculate the t-value,
2. use a t-table, and
3. interpret the results.

1. Calculating the T-Value

To calculate the t-value, you need to complete the following tasks:

(a) **Calculate the difference between the means.**
In the above example the difference is 156 - 154 = 2 cm.

(b) **Calculate the standard deviation** for each of the study groups. (The concept of standard deviation and how it is calculated has already been discussed in **Module 25**). Suppose the standard deviations shown in the final column of Table 27.1 were found.

(c) **Calculate the standard error of the difference between the two means.**

The standard error of the difference is given by the following formula:

$$\sqrt{\frac{SD_1^2}{n_1} + \frac{SD_2^2}{n_2}}$$

Where: SD_1 is the standard deviation of the first sample,
SD_2 is the standard deviation of the second sample,
n_1 is the sample size of the first sample,
n_2 is the sample size of the second sample.

For our data if we take the women with normal deliveries as sample 1 and those with Caesarean sections as sample 2 the standard error of the difference is:

$$\sqrt{\frac{3.1^2}{60} + \frac{2.8^2}{52}} = 0.56$$

(d) Finally, **divide the difference between the means by the standard error** of the difference. The value now obtained is called the **t-value.**

In the above example: $t = \frac{2}{0.56} = 3.6$

Expressed in one single formula:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{SD_1^2}{n_1} + \frac{SD_2^2}{n_2}}}$$

where $\bar{X}_1$ is the mean value of the first sample, and $\bar{X}_2$ is the mean value of the second sample.

2. Using a T-Table

Once the t-value has been calculated, you will have to refer to a t-table, from which you can determine whether the null hypothesis is rejected or not. **Annex 27.1** contains a t-table.

(a) First, decide which **significance level (p-value)** you want to use. (See **Module 26**). Remember that the p-value is an expression of the likelihood of finding a difference by chance when there is no real difference. Usually we take a p-value of 0.05.

(b) Second, determine the number of **degrees of freedom** for the test being performed. Degrees of freedom is a measure derived from the sample size, which has to be taken into account when performing a t-test. The bigger the sample size (and the degrees of freedom) the smaller the difference needed to reject the null hypothesis.

The way the number of degrees of freedom is calculated differs from one statistical test to another. For **Student's t-test** the number of degrees of freedom is calculated as the sum of the two sample sizes minus 2.

Thus, for **Example 1** the number of degrees of freedom is:

$$d.f. = 60 + 52 - 2 = 110.$$

Note: This is an approximate way of determining the number of degrees of freedom. For the exact method, refer to a statistics text.

(c) Third, the t-value belonging to the p-value and the degrees of freedom is located in the table.

In our example, we look up the t-value belonging to $p = 0.05$ and $d.f. = 120$ and we find it is 1.98.

3. Interpreting the Result

We now compare the absolute value of the t-value calculated in step 1 (i.e., the t-value, ignoring the sign) with the t-value derived from the table in step 2. If the calculated t-value is **larger** than the value derived from the table, p is **smaller** than the value indicated at the top of the column. We then **reject** the **null hypothesis** and conclude that there is a statistically significant difference between the two means.

If the calculated t-value is **smaller** than the value derived from the table, p is **larger** than the value indicated at the top of the table. We then **accept** the **null hypothesis** and conclude that the observed difference is not statistically significant.

In our example the t-value calculated in step 1 is 3.6, which is larger than the t-value derived from the table in step 2 (1.98). Thus, p is smaller than 0.05, and we therefore reject the null hypothesis and conclude that the observed difference of 2 cm between the mean heights of women with normal deliveries and women with Caesarean sections is a statistically significant difference.

We can express this conclusion in different ways:

- We can say that the probability that the observed difference of 2 cm between the two groups is due to chance is less than 5%.
- We can also say that the difference between the two groups is 3.6 times the standard error.

If you want to compare mean values of more than two groups (e.g., heights of urban, semi-urban, and rural women) you cannot use **Student's t-test**. In this case you must use the **F-test**, which is not described here.

III. CHI-SQUARE (χ^2) TEST

If you have categorical data the chi-square test is used to find out whether observed differences between proportions of events in groups may be considered statistically significant.

Example 2:

Suppose that in a study of the factors affecting the utilization of antenatal clinics you found that 64% of the women who lived within 10 kilometres of the clinic came for antenatal care, compared to only 47% of those who lived more than 10 kilometres away. This suggests that antenatal care (ANC) is used more often by women who live close to the clinics. The complete results are presented below:

Table 27.2. Utilization of antenatal clinics by women living far from and near the clinic.

Distance from ANC	Used ANC	Did not use ANC	Total
Less than 10 km	51 (64%)	29 (36%)	80 (100%)
10 km or more	35 (47%)	40 (53%)	75 (100%)
Total	86	69	155

From the table we determine that there seems to be a difference in utilization of antenatal care between those who live close to and those who live far from the clinic (64% versus 47%). We now want to know whether this observed difference is statistically significant.

The chi-square test can be used to give us the answer. This test is based on measuring the difference between the observed frequencies and the expected frequencies **if the null hypothesis (i.e., the hypothesis of no difference) were true.**

To perform a χ^2 test you need to complete the following 3 steps:

1. calculate the χ^2 value,
2. use a χ^2 table, and
3. interpret the result.

1. Calculate the χ^2 Value

Complete the following steps:

(a) Calculate the expected frequency (E) for each cell.

To find the expected frequency E of a cell you multiply the row total by the column total and divide by the grand (overall) total:

$$E = \frac{\text{row total x column total}}{\text{grand (overall) total}}$$

(b) For each cell, subtract the expected frequency from the observed frequency (O).
O - E

(c) For each cell, square the result of (O - E) and divide by the expected frequency E.

(d) Add the results of step (c) for all the cells.

The formula for calculating a chi-square value (steps (b) to (d)) is as follows:

$$\chi^2 = \Sigma \frac{(O - E)^2}{E}$$

where: O is the observed frequency (indicated in the table)
E is the expected frequency (to be calculated), and
Σ (the sum of) directs you to add together the products of $(O - E)^2$ for all the cells of the table.

For a two-by-two table (which contains 4 cells) the formula is:

$$\chi^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E2)^2}{E_2} + \frac{(O_3 - E_3)^2}{E_3} + \frac{(O_4 - E_4)^2}{E_4}$$

2. Using a χ^2 Table

As for the t-test, the calculated χ^2 value has to be compared with a theoretical χ^2 value in order to determine whether the null hypothesis is rejected or not. **Annex 27.2** contains a table of theoretical χ^2 values.

(a) First you must decide on a p-value. We usually take 0.05.

(b) Then the degrees of freedom have to be calculated. With the χ^2 test the number of degrees of freedom is related to the number of cells, i.e., the number of groups or variables you are comparing. The number of degrees of freedom is found by multiplying the number of rows (r) minus 1 by the number of columns (c) minus 1:

$$\text{d.f.} = (r - 1) \times (c - 1)$$

For a simple two-by-two table the number of degrees of freedom is 1 (i.e., d.f. = (2 - 1) x (2 - 1) = 1).

(c) Then the χ^2 value belonging to the p-value and the number of degrees of freedom is located in the table, in order to determine whether the χ^2 value is statistically significant or not.

3. Interpreting the Result

As for the t-test, the null hypothesis is rejected if $p < 0.05$, which is the case if the calculated χ^2 value is larger than the theoretical χ^2 value in the table.

Let us now apply the χ^2 test to the data given in Example 2 (utilization of antenatal care). This gives the following result:

Step 1(a):

The expected frequencies for each cell are calculated as follows:

$E_1 = 86 \times 80 / 155 = 44.4$ $\quad$ $E_2 = 69 \times 80 / 155 = 35.6$

$E_3 = 86 \times 75 / 155 = 41.6$ $\quad$ $E_4 = 69 \times 75 / 155 = 33.4$

For convenience, the observed and expected frequencies are shown in the following table:

Table 27.3. Utilization of antenatal clinics, observed and expected frequencies.

Distance from ANC	Used ANC	Did not use ANC	Total
Less than 10 km	$O_1 = 51$ $E_1 = 44.4$	$O_2 = 29$ $E_2 = 35.6$	80
10 km or more	$O_3 = 35$ $E_3 = 41.6$	$O_4 = 40$ $E_4 = 33.4$	75
Total	86	69	155

Note that the expected frequencies refer to the values we would have expected, given the total numbers of 80 and 75 women in the two groups, **if the null hypothesis**, stating that there is **no** difference between the two groups, **were true**.

Steps 1(b) to 1(d):

$$\chi^2 = \frac{(51-44.4)^2}{44.4} + \frac{(29-35.6)^2}{35.6} + \frac{(35-41.6)^2}{41.6} + \frac{(40-33.4)^2}{33.4}$$

$$= 0.98 + 1.22 + 1.05 + 1.30 = 4.55$$

Step 2:

As we have a simple two-by-two table the number of degrees of freedom (d.f.) is 1.

Use the table of chi-square values in Annex 27.2. We have decided beforehand on a level of significance of 5% (p-value = 0.05).

As the number of d.f. is 1, we look along that row in the column where p = 0.05. This gives us the value of 3.84. Our value of 4.55 is **larger** than 3.84, which means that the p value is **smaller** than 0.05 (it is even smaller than 0.01).

Step 3:

We can now conclude that the women living within a distance of 10 km from the clinic use antenatal care significantly more often than the women living more than 10 km away.

It is important to present your data clearly and to formulate carefully any conclusions based on statistical tests in the final report of your study.

For the above **example**, you could present Table 2 in the report and state your conclusions in this way:

> "Table 2 indicates that 64% of the women living within a distance of 10 km from the clinic used ante-natal care during pregnancy, compared to only 47% of women living 10 km or further away from the nearest clinic. This difference is statistically significant ($\chi^2 = 4.55$; $p < 0.05$)."

Note:

- The χ^2 test can only be applied if the sample is large enough. The general rule is that the total sample should be at least 40 **and** the **expected** frequency in each of the cells should be at least 5. If this is not the case, Fisher's exact test should be used (see Chapter 9 of Swinscow's *Statistics at Square One* referenced in Annex 26.1). If the table is more than a two-by-two table, the expected frequency of 1 in 5 cells is allowed to be less than 5.
- Unlike the t-test, the χ^2 test can also be used to compare more than two groups. In that case a table with three or more rows and columns would be designed, rather than a two-by-two table.

In the above example one could decide to distinguish between three different distances: less than 5 km, 5 to 10 km and more than 10 km. The data would then be put in a two-by-three table. The number of degrees of freedom would be (3 - 1) x (2 - 1) = 2.

Quick formula

For two-by-two tables there is a quick method for calculating the chi-square value, which can replace step 1 described above.

If the various numbers in the cross-table are represented by the following letters:

	Condition +	-	Total
Exposure			
Yes	a	b	e
No	c	d	f
Total	g	h	n

The quick formula for calculating the Chi-square value is

$$\chi^2 = \frac{n(ad - bc)^2}{efgh}$$

GROUP WORK

If your data was collected by unpaired observations, identify the appropriate significance test and perform the necessary analysis.

Annex 27.1. Student's t Distribution

The first column lists the number of degrees of freedom. The headings of the other columns give probabilities (P) for t to exceed the entry value. Use symmetry for negative values.

Degrees of freedom	t-value if p = 0.05	t-value if p = 0.01
1	12.71	63.66
2	4.30	9.92
3	3.18	5.84
4	2.78	4.60
5	2.57	4.03
6	2.45	3.71
7	2.36	3.50
8	2.31	3.36
9	2.26	3.25
10	2.23	3.17
11	2.20	3.11
12	2.18	3.05
13	2.16	3.01
14	2.14	2.98
15	2.13	2.95
16	2.12	2.92
17	2.11	2.90
18	2.10	2.88
19	2.09	2.86
20	2.09	2.85
21	2.08	2.83
22	2.07	2.82
23	2.07	2.81
24	2.06	2.80
25	2.06	2.79
30	2.04	2.76
40	2.02	2.70
60	2.00	2.66
120	1.98	2.62
Infinite	1.96	2.58

If the calculated t-value (ignoring the sign) is **larger** than the value indicated in the table, the p-value is **smaller** than that indicated at the top of the column.

In that case, the null hypothesis, stating that there is no difference, is **rejected**, and it can be concluded that there **is** a significant difference.

Annex 27.2. Table of χ^2 Values

Degrees of freedom	χ^2-value if p = 0.05	χ^2-value if p = 0.01
1	3.84	6.63
2	5.99	9.21
3	7.81	11.34
4	9.49	13.28
5	11.07	15.09
6	12.59	16.81
7	14.07	18.48
8	15.51	20.09
9	16.92	21.67
10	18.31	23.21
11	19.68	24.72
12	21.03	26.22

If the calculated χ^2 value is **larger** than the value indicated in the table, the p-value is **smaller** than that indicated at the top of the column.

In that case, the null hypothesis, stating that there is no difference, is **rejected**, and it can be concluded that the difference **is** statistically significant.

Annex 27.3. Dealing with Confounding Variables: Mantel-Haenszel Chi-Squared Test

In Table 27.4 the results of a schistosomiasis survey among the inhabitants of two villages are presented.

Table 27.4. Prevalence of schistosomiasis in two villages, A and B.

	Village A	Village B	Total
Schisto (+)	80 (32%)	80 (32%)	160
No schisto (-)	170 (68%)	170 (68%)	340
Total	250 (100%)	250 (100%)	500

It seems that the prevalence of schistosomiasis is the same in both villages (32%).

However, the researchers suspect that age is a confounding variable. Therefore, Table 27.4 is split up into two tables (27.5 and 27.6). Note that adding the numbers in Tables 27.5 and 27.6 will give us Table 27.4.

Table 27.5. Prevalence of schistosomiasis in children aged 5-19 in Villages A and B.

	Village A	Village B	Total
Schisto (+)	37 (62%)	73 (38%)	110
No schisto (-)	23 (38%)	117 (62%)	140
Total	60 (100%)	190 (100%)	250

$\chi^2 = 9.08$; 1 degree of freedom; $p < 0.01$.

Table 27.6. Prevalence of schistosomiasis in those aged 20 years and above in Villages A and B.

	Village A	Village B	Total
Schisto (+)	43 (23%)	7 (12%)	50
No schisto (-)	147 (77%)	53 (88%)	200
Total	190 (100%)	60 (100%)	250

$\chi^2 = 2.78$; 1 degree of freedom; $p > 0.05$.

From Tables 27.5 and 27.6 it becomes clear that:

- Within each age group schistosomiasis is more prevalent in Village A than in Village B.
- Schistosomiasis is more prevalent in children than in adults.

- In Village A there are relatively few children and many adults compared to Village B.

Age is said to be a confounding variable because it is related to the variable of interest (prevalence of schistosomiasis) and to the groups being compared (residence in Village A or B).

This example illustrates a very important point in analyzing data. It may be very misleading to pool dissimilar data. In this particular example, pooling the age groups masked an important real difference. In other situations pooling the data may suggest a difference or association that does not exist or even a difference opposite to that which exists.

It is, therefore, important to analyze the above data for the different age groups separately. The appropriate χ^2 values (with continuity correction) for comparing the prevalences in Villages A and B are shown in Tables 27.5 and 27.6. The difference in prevalence is significant for children but not for adults.

Mantel-Haenszel χ^2 test

It is often useful to have a summary test which pools the evidence from the individual tables, but takes into account the confounding factor (age in our example). The Mantel-Haenszel χ^2 test for doing this will be described.

For each of the two-by-two tables we will use the notation:

a	b	e
c	d	f
g	h	n

The following three values are calculated from each table, then summed over the tables:

1. the observed value for a, O_a
2. the expected value for a, E_a, which equals eg/n
3. the variance of a, V_a, which equals $efgh/(n^2(n-1))$
   The χ^2 value with continuity correction is

$$\chi^2 = \frac{(O_a - E_a - 0.5)^2}{V_a} \qquad \text{with degrees of freedom} = 1$$

In the **example**, the calculations are

	O_a	$E_a = eg/n$	$V_a = efgh/(n^2(n-1))$
Children	37	26.4	$110 \times 140 \times 60 \times 190 / 250^2 \times 249 = 11.3$
Adults	43	38	$50 \times 200 \times 190 \times 60 / 250^2 \times 249 = 7.3$
Total	80	64.4	18.6

$$\chi^2 = \frac{(80 - 64.4 - 0.5)^2}{18.6} = \frac{15.1^2}{18.6} = 12.25 \quad (p < 0.001)$$

It can, therefore, be concluded that the prevalence of schistosomiasis is significantly different in Village A and B. (Remember that this seemed not to be the case when we looked at Table 27.4, in which the data for both adults and children were pooled.)

Validity of Mantel-Haenszel χ^2 test

The Mantel-Haenszel χ^2 test is an approximate test. The rule for assessing its adequacy is more complicated than that for the ordinary χ^2 test. Two additional values are calculated for each table and summed over the tables. These are

1. min(e, g), that is the smaller of e and g
2. max(0, g - f), that is 0 if g is smaller than or equal to f and g - f if g is larger.

Both these sums should differ from the total of the expected values, E_a, by at least 5. The details of the calculation for the above example are:

	min (e, g)	max (0, g - f)
Children	60	0
Adults	50	0
Total	110	0

These sums are 110 and 0, both of which differ from 64.4 (E_a) by more than 5. The use of the Mantel-Haenszel test is therefore valid.

Trainer's Notes

Module 27: DETERMINING DIFFERENCES BETWEEN GROUPS: PART I — ANALYSIS OF UNPAIRED OBSERVATIONS

Timing and Teaching Methods

1 hour Introduction and discussion
3 hours* Group work

Introduction and Discussion

- We advise that you present either the t-test or the χ^2 test immediately after completing **Module 26**, so that participants get a better idea of what a significance test is and how it is used. Probably all research teams will be applying χ^2 tests on their data, but not all teams will be using the t-test. Therefore, you might present the χ^2 test in combination with **Module 26** and the t-test in another session.

- Proceed slowly, step-by-step, when explaining how significance tests work, so as not to frighten participants who have little experience. Stress that it is not important to understand WHY the calculations of t-values and χ^2-values are performed this way (actually there is not always a clear rationale, for example, for the concept of degrees of freedom). It is enough to know HOW they are done. Be careful with formulae: present them only after the step-wise calculations have been explained.

- If you wish, you may use examples taken from the groups' own studies rather than Examples 1 and 2 presented in the module. Remember, however, to use simple examples, i.e. two-by-two tables and small numbers, so that they are easier to follow.

- Take extra care to explain how to use the t-table and χ^2-table and how to interpret the results. Let participants struggle with the examples themselves before providing the correct answers.

- Also pay attention to the appropriate phrasing of conclusions based on significance tests, both in cases where the results are significant and where they are not significant.

- **Annex 27.3** should not be presented unless the group is very advanced. The teams can read it and use the test, if necessary.

Group Work

When performing statistical tests make sure that each member of the group does at least one test on his or her own.

Health Systems Research Training Series

Volume 2, Part II: Data Analysis and Report Writing

Module 28:

DETERMINING DIFFERENCES BETWEEN GROUPS: PART II — ANALYSIS OF PAIRED OBSERVATIONS

STEPS IN DATA ANALYSIS AND REPORT WRITING

Questions you must ask	Steps you will take*	Important elements of each step
What data have been collected for each research objective? Are data complete, accurate?	Prepare data for analysis	Review field experiences Inventory data for each objective/study population Sort data and check quality Check computer outputs
What do the data look like?	Describe variables	Frequency distributions Figures, means
How can the data be summarized for easy analysis?	Cross-tabulate quantitative data Summarize qualitative data	Cross-tabulate in relation to objectives Graphic displays, narratives
For quantitative data: does each research objective aim to describe, compare, or find associations?	Determine the type of statistical analysis required	Review objectives, study type, and variables Statistical description of variables Choosing significance tests
1. How can the data be described?		
2. How can differences between groups be determined?	Analyze paired and unpaired observations	**Student's t-test **Paired t-test **Chi-square test **McNemar's chi-square test
3. How can the associations between variables be determined?	Measure associations between variables	**Scatter diagram **Regression line and correlation coefficient **Relative risk, odds ratio
How should the report be written?	Write the report and formulate recommendations	Prepare outline for report Draft and redraft Summarize findings Summarize conclusions for each objective Formulate recommendations Prepare abstract
How should the findings and recommendations be presented and disseminated?	Present a summary and draft a plan for implementation of recommendations	Discuss summaries with different target groups Discuss plan for implementation

* These steps need not be in the sequence in this diagram. The sequence may be adjusted according to the needs of the workshop.
** These elements are optional and may be omitted if not relevant in a particular course.

Module 28: DETERMINING DIFFERENCES BETWEEN GROUPS: PART II — ANALYSIS OF PAIRED OBSERVATIONS

OBJECTIVES:

At the end of this session you should be able to:

1. Identify research studies where pairing or matching of subjects is necessary.
2. Identify and use the significance tests appropriate for studies using paired data.

I. **Introduction**

II. **Paired T-Test**

III. **McNemar's Chi-Squared Test**

I. INTRODUCTION

In **Module 23** the concept of paired or matched observations was discussed. The example presented in that module dealt with nominal data. However, paired observations can also be made when the data are numerical. This module describes the most commonly used tests for paired observations:

- the **paired t-test** for numerical data and
- **McNemar's chi-squared test** for nominal data.

What is meant by pairing or matching?

The concept of pairing or matching subjects is illustrated by the following examples:

Example 1:

A researcher wanted to find out whether a class of students taught with audiovisual aids receive, on average, better grades than those who are taught without audiovisual (AV) aids. To minimize the effect of confounding variables such as social status, previous knowledge of the subjects, and IQ levels, each student in the AV class was paired with another in the non-AV class for these variables.

Example 2:

During a nutritional survey, a quality-control exercise was carried out to check the agreement between two observers in measuring children's weight. In this instance each child acted as his own match.

Example 3:

A study team compared schistosomiasis egg counts in two villages. It recognized that egg counts vary with age and sex. It decided to ensure that the samples were comparable with respect to age and sex by selecting subjects in pairs, with one member of each pair from each village, who were matched for age and sex.

II. PAIRED T-TEST

In **Module 27** a comparison of sample means was performed for unpaired numerical observations by using the **t-test**. When dealing with paired (matched) observations, comparison of sample means is performed by a modified t-test known as the **paired t-test**.

In the **paired t-test** a single set of differences between the paired observations is used instead of the original two sets of observations.

The **paired t-test** calculates a value of t as:

$$t = \frac{\text{mean difference}}{\text{standard error}}$$

The number of degrees of freedom is the sample size minus 1 (or the number of paired observations minus 1).

To interpret a result, the same table of t-values is used as for the **t-test** that is used for unpaired observations (see **Annex 27.1**).

To illustrate how the **paired t-test** is used, it will be performed on the results of the nutritional survey referred to in **Example 2** above. The results are:

Table 28.1. Results of quality control exercise during a nutrition survey.

Child no.	Weight measurement (kg)		Difference A - B (kg)
	Observer A	Observer B	
1	18.6	17.7	0.9
2	17.1	14.5	2.6
3	14.3	12.4	1.9
4	23.2	20.7	2.5
5	18.4	16.8	1.6
6	14.9	14.4	0.5
7	16.6	14.1	2.5
8	14.8	17.1	-2.3
9	21.5	21.2	0.3
10	24.6	21.9	2.7
11	17.4	16.6	0.8
12	15.7	13.6	2.1
13	16.1	14.5	1.6
14	12.9	11.2	1.7
15	12.3	16.0	-3.7
16	19.4	20.4	-1.0
17	19.3	17.5	1.8
18	24.8	22.2	2.6
19	14.3	15.1	-0.8
20	13.4	10.9	2.5

The null hypothesis in this study is that if observers A and B measured all the children in the population from which these 20 children were sampled, there would, on average, be no difference between their measurements. In other words, the mean difference between A and B would be zero.

We can regard this set of 20 differences (the A - B column) as a sample from the population of differences that would have been obtained if the observers had measured the whole population.

To perform the significance test the value of t has to be calculated and compared to the theoretical value in the t-table to determine the probability that the result occurred by chance.

This is done as follows:

1. Calculate the mean difference in the sample. This is the sum of the differences divided by the number of measurements:

   Mean difference = 1.04

2. Calculate the standard deviation of the differences (**Module 25**):

Standard deviation = 1.77

3. Calculate the standard error (**Module 25**):

$$\text{Standard error} = \frac{\text{Standard deviation}}{\sqrt{\text{sample size}}} = \frac{1.77}{\sqrt{20}} = 0.40$$

4. The value of t is the mean difference divided by the standard error:

$$t = \frac{1.04}{0.40} = 2.60$$

5. Refer to the table of t-values in **Annex 27.1**.

   The number of degrees of freedom is the sample size (the number of pairs of observations) minus 1, which in this case is 20 - 1 = 19.

   The probability from the table is < 0.05 which allows us to conclude that there is a significant difference between the observers.

III. MCNEMAR'S CHI-SQUARED TEST

The **McNemar's chi-squared test** is used with NOMINAL data to compare PROPORTIONS of paired observations. It is important to note that the layout of the table is different from that used with unpaired samples.

Table 28.2 shows the results of a case-control study that was conducted to determine causes of a cholera outbreak in Italy. For each cholera case confirmed in the hospital a subject was sought of the same sex, the same age decade, and the same neighbourhood.

However, the layout of **Table 28.2** is not correct, as it does not take account of the fact that cases and controls were selected as pairs.

The correct layout is presented in **Table 28.3**.

Table 28.2. Consumption of fish or seafood by cholera patients and controls in 5 days preceding illness (incorrect layout).

	Cases	Controls	Total
Ate seafood	42 (55%)	15 (20%)	57
Did not eat seafood	34 (45%)	61 (80%)	95
Total	76 (100%)	79 (100%)	152

Table 28.3. Consumption of fish or seafood by cholera patient/control pairs in 5 days preceding illness (correct layout).

	Cases		
Controls	Ate seafood	Did not eat seafood	Total
Ate seafood	12 (16%)	3 (4%)	15
Did not eat seafood	30 (39%)	31 (41%)	61
Total	42	34	76 (100%)

Source: Baine, W.B., Mazzotti, M, Greco, D, Izzo, E., Zampieri, A., Angioni, G., Di Gioia, M., Gangarosa, E.J., Pocchiari, S. 1974. Epidemiology of cholera in Italy in 1973. Lancet, ii (Dec.), 1370-1374.

How should we interpret **Table 28.3**?

In 12 pairs both cases and controls ate seafood and in 31 pairs neither of them ate seafood. These 43 pairs, therefore, give us no information about whether eating seafood is a risk factor for getting cholera. However, in 30 pairs (39%) the cases ate seafood, while the controls did not, whereas in only 3 pairs (4%) the controls ate seafood while the cases did not. It would seem, therefore, that eating seafood was a risk factor for getting cholera.

Before we accept that conclusion, we must perform a significance test to estimate the likelihood that these results are due to chance or sampling variation only. In this case the appropriate significance test is **McNemar's chi-squared test**:

$$\chi^2 = \frac{(|r - s| - 1)^2}{r + s} \quad \text{with 1 degree of freedom}$$

where: r is the number of +- responses,
s the number of -+ responses, and
$|r - s|$ is the difference between r and s as a positive number, irrespective of whether r or s is the larger number.

Note:

McNemar's χ^2 test is only valid if $(r + s)$ is larger than 10.

The test can be performed on the data in our example because $r + s$ $(30 + 3)$ is larger than 10.

The calculation of the chi-square value is as follows:

$$\chi^2 = \frac{(30 - 3 - 1)^2}{30 + 3} = \frac{26^2}{33} = 20.5 \quad \text{with 1 degree of freedom.}$$

From the table of χ^2 values (**Annex 27.2**) we see that $p < 0.01$. This means that if there were no real association between eating seafood and getting cholera (null hypothesis) the chance of getting a result

as discrepant as the one shown in **Table 28.3** would be less than 1 in 100. We therefore reject the null hypothesis and conclude that eating seafood was a risk factor for getting cholera.[1]

GROUP WORK

If your data were collected by paired or matched observations, identify the appropriate statistical test and make the necessary calculations and analysis.

[1] Part of this module stems from the course material for the Masters of Science degree at the London School of Hygiene and Tropical Medicine.

Trainer's Notes

Module 28: DETERMINING DIFFERENCES BETWEEN GROUPS: ANALYSIS OF PAIRED OBSERVATIONS

Timing and Teaching Methods

1/2 hour	Introduction and discussion
2 hours	Group work

Introduction and Discussion

- If none of the teams has paired observations and if participants have little experience with statistics, this module need not be presented.
- When explaining how to calculate the t-values and χ^2-values proceed step-by-step very slowly. Again, it is more important that participants understand **how** to do the calculations, rather than why they are done this way.
- Take time to ensure that everyone knows how to read **Table 28.3**. This is the first table so far in which the numbers represent **pairs** of observations.
- Make sure that participants know how to use the t-table and χ^2-table and how to interpret the results.

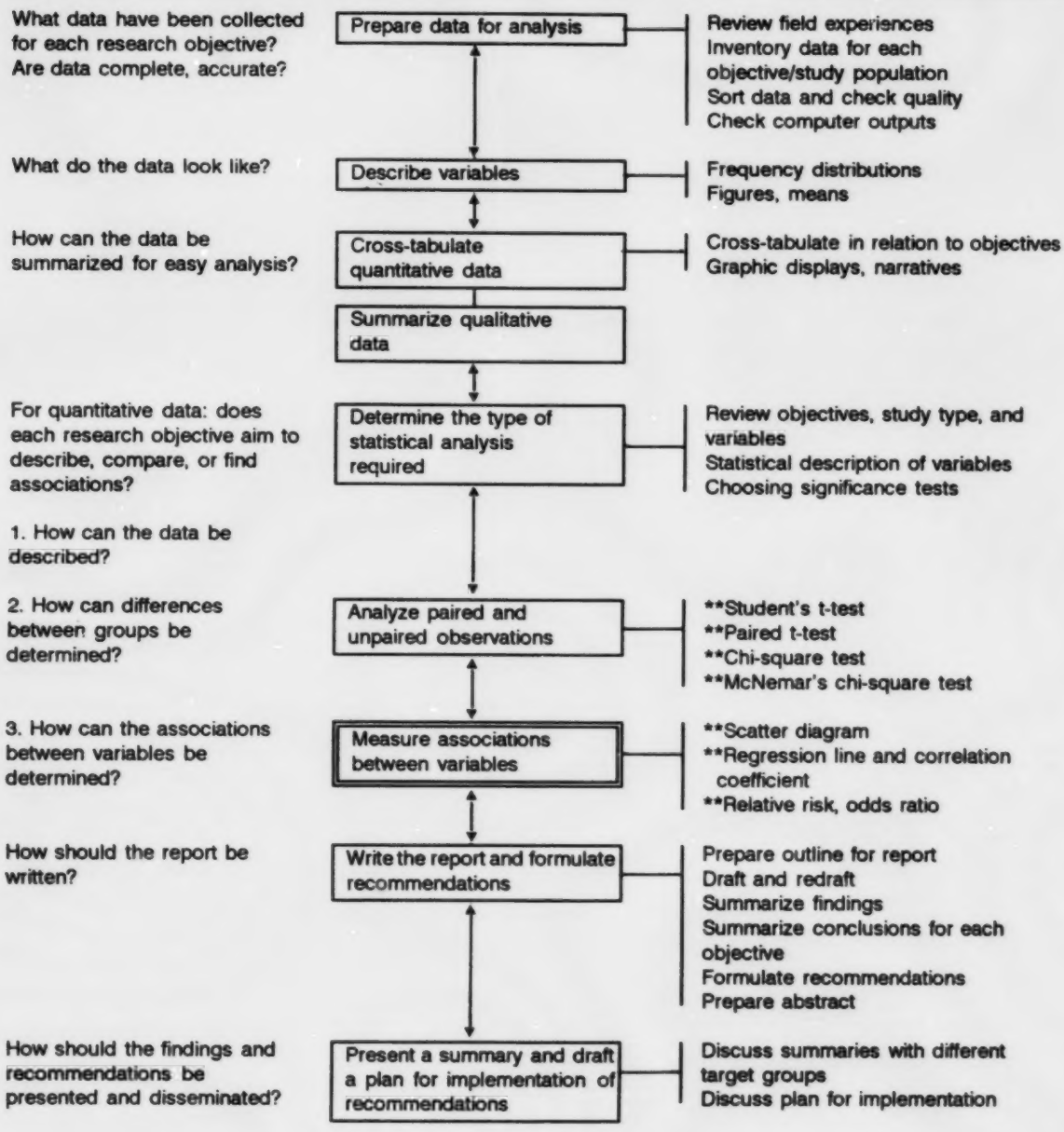
Health Systems Research Training Series

Volume 2, Part II: Data Analysis and Report Writing

Module 29:

MEASURING ASSOCIATIONS BETWEEN VARIABLES: REGRESSION AND CORRELATION

STEPS IN DATA ANALYSIS AND REPORT WRITING


Questions you must ask
Steps you will take*
Important elements of each step
What data have been collected for each research objective? Are data complete, accurate?
Prepare data for analysis
Review field experiences
Inventory data for each objective/study population
Sort data and check quality
Check computer outputs
What do the data look like?
Describe variables
Frequency distributions
Figures, means
How can the data be summarized for easy analysis?
Cross-tabulate quantitative data
Cross-tabulate in relation to objectives
Graphic displays, narratives
Summarize qualitative data
For quantitative data: does each research objective aim to describe, compare, or find associations?
Determine the type of statistical analysis required
Review objectives, study type, and variables
Statistical description of variables
Choosing significance tests
1. How can the data be described?
2. How can differences between groups be determined?
Analyze paired and unpaired observations
**Student's t-test
**Paired t-test
**Chi-square test
**McNemar's chi-square test
3. How can the associations between variables be determined?
Measure associations between variables
**Scatter diagram
**Regression line and correlation coefficient
**Relative risk, odds ratio
How should the report be written?
Write the report and formulate recommendations
Prepare outline for report
Draft and redraft
Summarize findings
Summarize conclusions for each objective
Formulate recommendations
Prepare abstract
How should the findings and recommendations be presented and disseminated?
Present a summary and draft a plan for implementation of recommendations
Discuss summaries with different target groups
Discuss plan for implementation

* These steps need not be in the sequence in this diagram. The sequence may be adjusted according to the needs of the workshop.

** These elements are optional and may be omitted if not relevant in a particular course.

Module 29: MEASURING ASSOCIATIONS BETWEEN VARIABLES: REGRESSION AND CORRELATION

OBJECTIVES

At the end of this session you should be able to:

1. Illustrate the relationship between two numerical variables in a scatter diagram.
2. Interpret a regression line.
3. Calculate and interpret a correlation coefficient.
4. Perform a test for the significance of the correlation coefficient.

I. Introduction

II. Scatter Diagrams

III. Determining Linear Relationship: Fitting a Regression Line

IV. Correlation Coefficients

V. Testing the Significance of a Correlation Coefficient

I. INTRODUCTION

When exploring associations between variables we have to distinguish between nominal data, ordinal data, and numerical data (refer to Table 26.1 in **Module 26**).

Module 30 will deal with associations between nominal data, concentrating on case-control studies.

For associations between ordinal data, in which case **Spearman's rank correlation coefficient** or **Kendall's tau** can be calculated and tested for significance, you may refer to a textbook on statistics (see **Annex 26.1**).

In this module we will examine associations between numerical data where a linear relationship is suspected.

II. SCATTER DIAGRAMS

The first step in examining the relationship between two numerical variables, measured on the same subjects, is always to draw a **SCATTER DIAGRAM.**

Example 1:

In a nutrition study in a large rural district, a sample of 20 children 5 years of age were weighed and their family incomes estimated. The results were as follows:

Table 29.1. Weights and family incomes of 20 children 5 years of age.

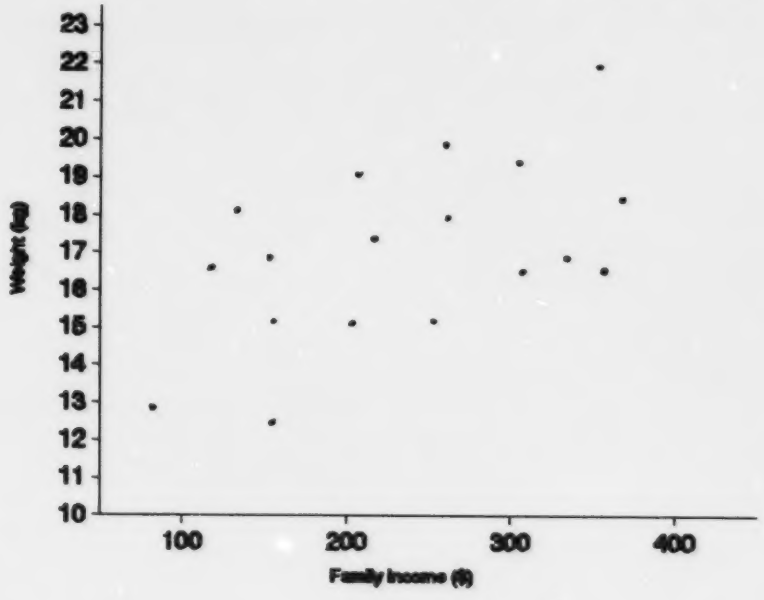
Family income in $/year	Weight in kg	Family income in $/year	Weight in kg
130	15.5	225	18.1
200	19.8	95	17.4
345	21.5	130	17.9
245	16.8	330	17.0
155	12.6	295	18.7
300	16.6	170	16.0
360	18.1	250	18.2
105	18.7	355	16.4
80	13.1	220	15.4
275	20.1	175	17.6

The objective was to examine whether, for this sample of children, weight and family income were related. It would be possible to divide the children into two income categories, a high-income category (e.g., $200 or more) and a low-income category (e.g., less than $200), and to calculate and compare the mean weight in each category to see if there is a difference. One would have to use a t-test to determine if the difference is significant. (See **Module 27**.)

After performing this analysis you might conclude that children from low-income families had lower weight, on average, than children from families with high incomes. However, it would be more interesting to take account of all the individual measurements and investigate whether the two variables "family income" and "weight of five year olds" are associated (i.e., if we can predict the value of one variable when the value of the other is known).

The following scatter diagram can be drawn:

Figure 29.1. Weights and family incomes of 20 children 5 years of age.



Notes on Drawing Scatter Diagrams

1. If we are examining how a dependent variable is associated with an independent variable, we generally put the dependent variable on the vertical axis (the y-axis) and the independent variable on the horizontal axis (the x-axis). Sometimes it is not clear which is the dependent variable, in which case the choice of axis is arbitrary.

2. Select the scales so that the scatter fills a reasonable portion of the diagram.

3. If an axis does not start from zero, show this clearly by "breaking" the axis (as has been done for the weight axis in the above example).

4. Label the axes clearly.

5. The plotted points should be big enough to stand out, so that the scatter is easy to look at.

III. DETERMINING LINEAR RELATIONSHIP: FITTING A REGRESSION LINE

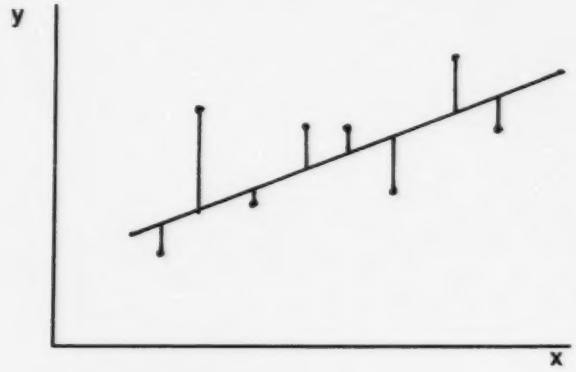
In the scatter diagram above there appears to be an upward trend in weight, with increasing family income. We can draw a line through the scatter of points, as a simple summary of the relationship between these two variables. This can be done "by eye" - a transparent ruler is useful for this. However, fitting by eye is rather subjective and we would prefer a more objective technique. The **LEAST SQUARES METHOD** gives the "best" line, using a technique that will be described below.

Any straight line drawn on a graph can be represented by the equation:

$$y = a + bx$$

Each point on the line has an x value and a y value, and the equation tells us how these x and y values are related. Different straight lines correspond to different values of a and b. The value of a tells us the INTERCEPT of the line on the y-axis (a is the distance to the point where the line crosses the y-axis from zero) and b tells us the SLOPE of the line.

To decide on the line to fit through our scatter, we have to decide what values of a and b to use. Basically, we choose them in such a way that the vertical distances of the points from the line are minimized. (To be more precise, we choose a and b that minimize the sum of the squares of these vertical distances. Hence the name "least squares method".)



Annex 29.1 explains how the values of a and b are calculated from the data. Some calculators and computer programs give the values of a and b automatically.

In our example, we find, using an appropriate calculator:

$a = 15.09$ $\quad$ $b = 0.00984$

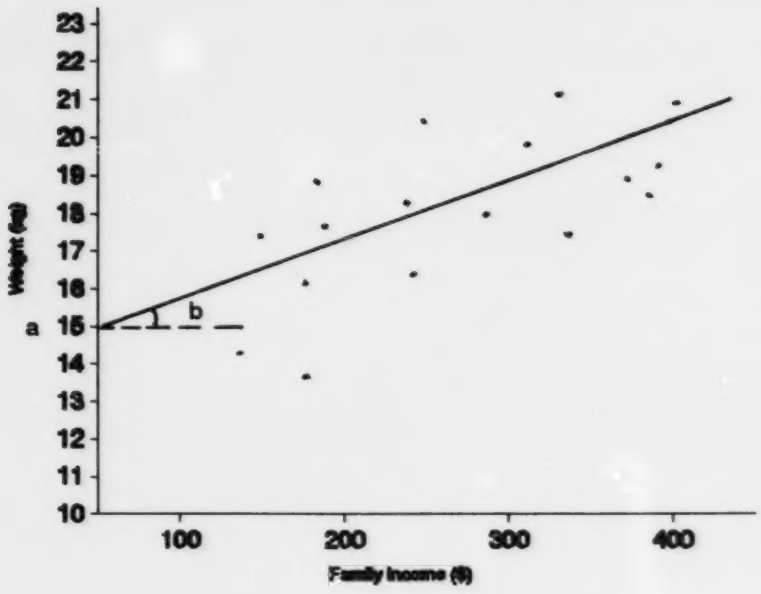
So the equation of our fitted line is

$$y = 15.09 + 0.00984x$$

To draw this on the scatter diagram we choose two values of x, find the corresponding values of y using the equation, plot the two points on the graph and join them with a straight line.

For example:

$x = 0$ gives $y = 15.09$
$x = 400$ gives $y = 15.09 + (0.00984 \times 400) = 19.03$



The fitted line is called the **LINEAR REGRESSION** of weight on family income.

Interpretation of a Regression Line

The regression line estimates the **average** value of y for a given value of x. For example, it tells us that children whose families have an income of $200/year would **on average** weigh about 17 kg, though some would weigh more and some less than this.

The slope b, called the REGRESSION COEFFICIENT, tells us the increase in the average value of y corresponding to a unit increase in x. So in our example, mean weight increased by 0.00984 kg (or about 10 g) for each increase of $1 in family income (or about 1 kg weight gain per $100 increase).

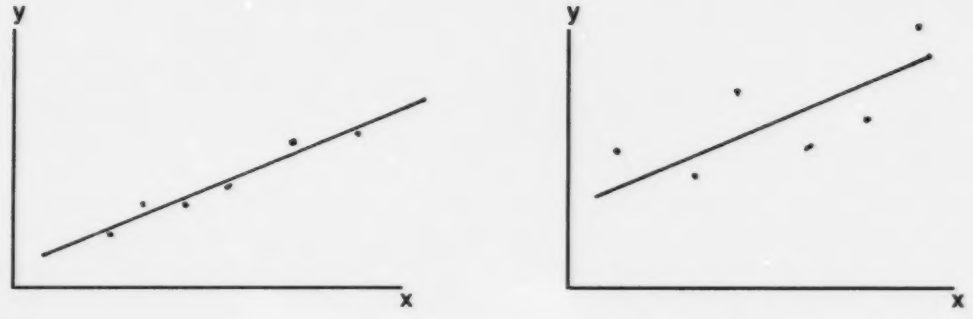
Two words of caution:

- A straight line should be fitted only if the scatter diagram suggests that the relationship between the two variables is roughly linear. More complex methods are available for fitting curves to the data.

- It is dangerous to **extrapolate** the regression line outside the range of the data. In our example, extrapolating the line to an income of $2000/year would yield an estimated mean weight of 34.8 kg, which is of course absurd.

IV. CORRELATION COEFFICIENTS

Consider the following two scatter diagrams:



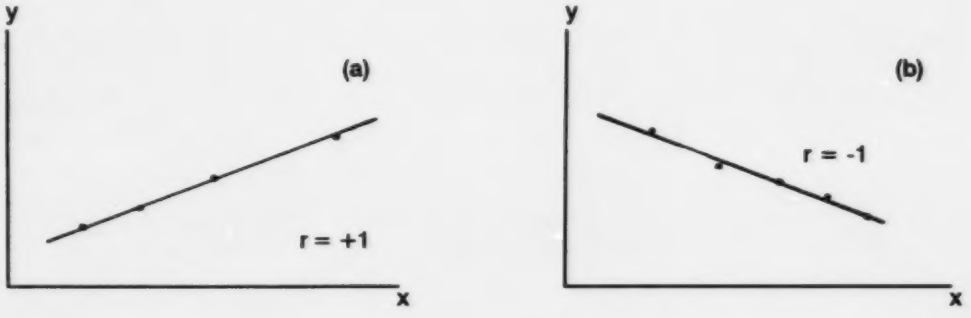
The regression coefficients b (i.e., the slopes of the lines) are identical in these two examples, but the scatter about the line is much greater in the second. Clearly the relationship between variables y and x is much closer in the first diagram.

The aim of PEARSON'S CORRELATION COEFFICIENT (r) is to measure the precision of the linear relationship between two variables.

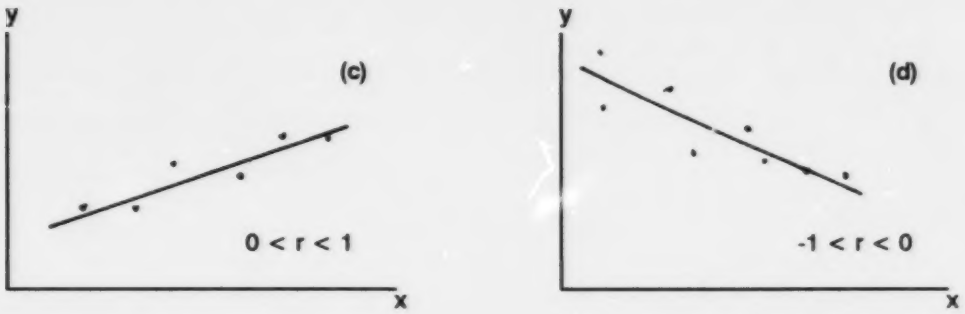
Some calculators and computer programs give the value of r automatically, at the same time as a and b. **Annex 30.1** shows how r is calculated. The correlation coefficient has the following properties:

1. For any data set, r lies between -1 and +1.

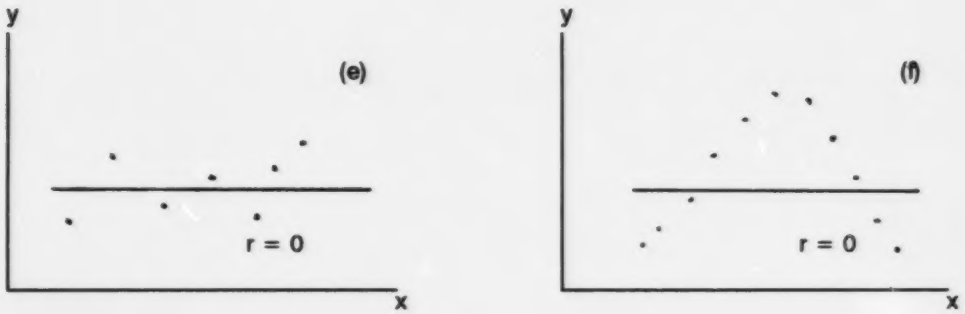
2. If r = +1, or -1, the relationship is perfect, that is, all the points lie exactly on the regression line. If r = +1, variable y increases as x increases (i.e., the line slopes upwards). If r = -1, variable y decreases as x increases (i.e., the line slopes downward).



3. If r lies between 0 and +1, the regression line slopes upwards, but the points are scattered about the line. The closer r is to 1, the closer the points are to the line. The same is true of negative values of r, between 0 and -1, but in this case the regression line slopes downward.



4. If r = 0, there is no **linear** relationship between y and x. This may mean that there is no relationship at all between the two variables (i.e., knowing x tells us nothing about the value of y). However, we could also obtain r = 0 if there were a **curved** relationship between y and x (see Diagram f).



5. A useful interpretation of r is that its square (r^2) measures the proportion of the variability in variable y **accounted for** by the linear relationship with variable x.

Returning to our example of weight and family income the calculator gives:

r = 0.414

which is positive (indicating an upward sloping line), but a long way from 1 (indicating plenty of scatter about the line).

V. TESTING THE SIGNIFICANCE OF A CORRELATION COEFFICIENT

The fitted regression line and the value of r were both calculated from a sample of just 20 children. The results are, therefore, subject to sampling error and are unlikely to be equal to the true regression line and the true value of r, which we would obtain if we measured all 5-year-old children in this district.

The question arises as to whether there is really any relationship at all between weight and income. Perhaps in the entire population of 5-year-old children the scatter diagram would look like diagram e above (no relationship between y and x) and the positive relationship in our sample occurred by chance.

To assess whether this is the case, we do a significance test on r. The null hypothesis is that in the whole population there is no linear relationship between y and x. To do the test we calculate

$$t = r \text{ times } \sqrt{\frac{n-2}{1-r^2}}$$

We compare this value of t to tables of the t distribution with (n - 2) degrees of freedom, where n is the number of observations.

In our example: $n = 20$ $r = 0.414$

Therefore $t = 0.414 \times \sqrt{\frac{18}{1 - 0.414^2}} = 1.93$

Since $t_{18,0.05} = 2.10$, the relationship is not significant ($p > 0.05$). This is actually a "borderline" case, because the p-value is a little higher than the conventional $p = 0.05$ cut-off point.

Association and Causation

Note that the existence of a statistical association, even if it is very strong, does NOT establish that an increase in x **causes** an increase in y, or that an increase in y causes an increase in x. A fundamental weakness of observational studies is that they can demonstrate association but not causation. To demonstrate a causal relationship one would need to choose an experimental study design.[2]

[2] Most of this module stems from the MSc course materials of the London School of Hygiene and Tropical Medicine.

Annex 29.1. Calculation of the Regression Coefficient and Correlation Coefficient

For the regression equation

$$y = a + bx$$

the letters a and b (the regression coefficient) are calculated as follows from the data:

$$b = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2} = \frac{\Sigma xy - (\Sigma x)(\Sigma y) / n}{\Sigma x^2 - (\Sigma x)^2 / n}$$

$$a = \bar{y} - b\bar{x}$$

n is the number of observations,
$\bar{x}$ is the mean of all x-values,
$\bar{y}$ is the mean of all y-values.

The calculation of the correlation coefficient is as follows:

$$r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\sqrt{\Sigma(x - \bar{x})^2 \, \Sigma(y - \bar{y})^2}} = \frac{\Sigma xy - (\Sigma x)(\Sigma y) / n}{\sqrt{(\Sigma x^2 - (\Sigma x)^2/n)\,(\Sigma y^2 - (\Sigma y)^2/n)}}$$

Trainer's Notes

Module 29: MEASURING ASSOCIATIONS BETWEEN VARIABLES: REGRESSION AND CORRELATION

Timing and Teaching Methods

1 hour	Introduction and discussion
3 hours	Group work

Introduction and Discussion

- This module should only be presented if at least one of the research teams needs it to analyze its data or if the participants have sufficient background in statistics.

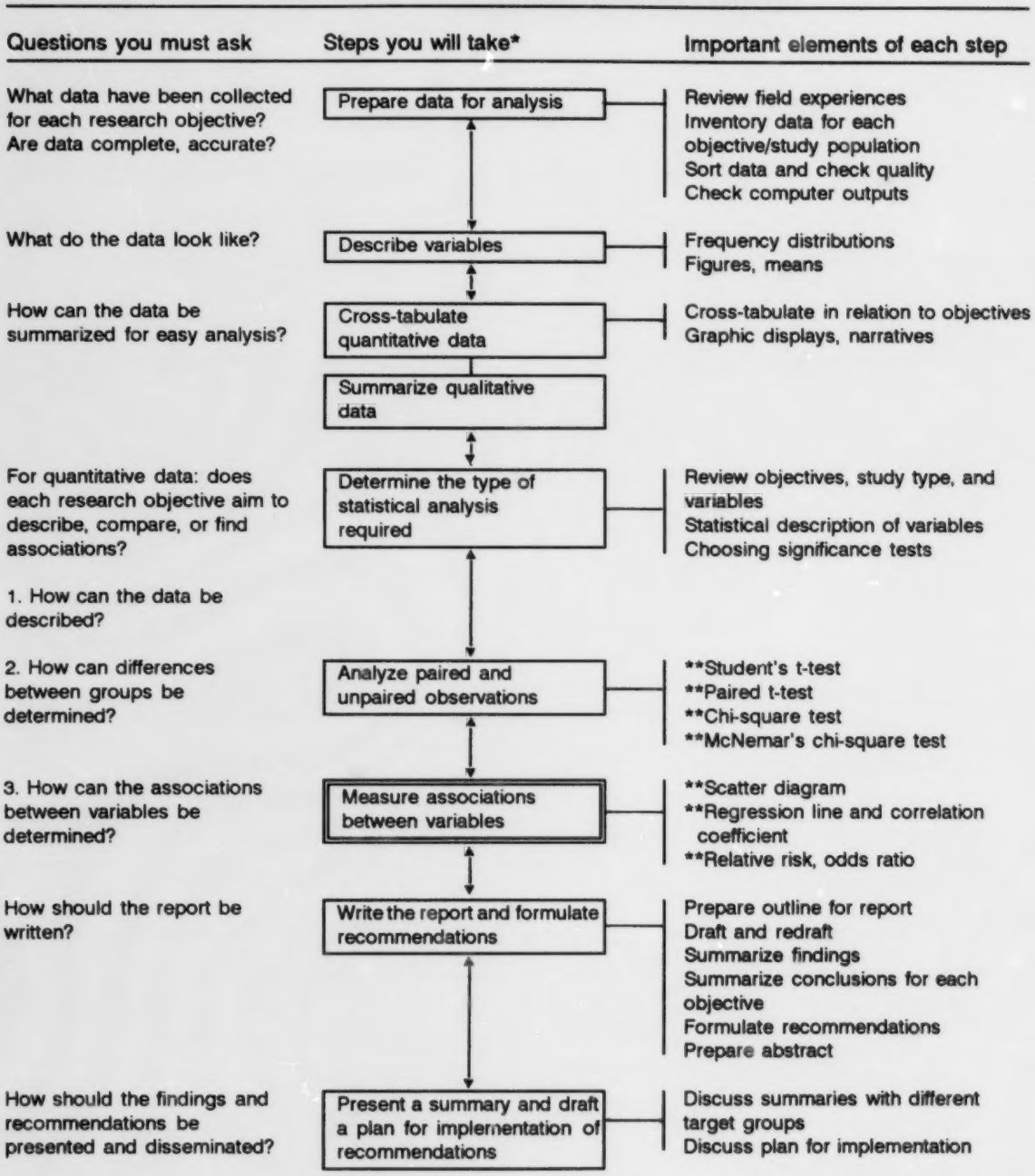
Health Systems Research Training Series

Volume 2, Part II: Data Analysis and Report Writing

Module 30:

MEASURES OF RISK IN CASE-CONTROL STUDIES

STEPS IN DATA ANALYSIS AND REPORT WRITING


Questions you must ask
Steps you will take*
Important elements of each step
What data have been collected for each research objective? Are data complete, accurate?
Prepare data for analysis
Review field experiences
Inventory data for each objective/study population
Sort data and check quality
Check computer outputs
What do the data look like?
Describe variables
Frequency distributions
Figures, means
How can the data be summarized for easy analysis?
Cross-tabulate quantitative data
Summarize qualitative data
Cross-tabulate in relation to objectives
Graphic displays, narratives
For quantitative data: does each research objective aim to describe, compare, or find associations?
Determine the type of statistical analysis required
Review objectives, study type, and variables
Statistical description of variables
Choosing significance tests
1. How can the data be described?
2. How can differences between groups be determined?
Analyze paired and unpaired observations
**Student's t-test
**Paired t-test
**Chi-square test
**McNemar's chi-square test
3. How can the associations between variables be determined?
Measure associations between variables
**Scatter diagram
**Regression line and correlation coefficient
**Relative risk, odds ratio
How should the report be written?
Write the report and formulate recommendations
Prepare outline for report
Draft and redraft
Summarize findings
Summarize conclusions for each objective
Formulate recommendations
Prepare abstract
How should the findings and recommendations be presented and disseminated?
Present a summary and draft a plan for implementation of recommendations
Discuss summaries with different target groups
Discuss plan for implementation

* These steps need not be in the sequence in this diagram. The sequence may be adjusted according to the needs of the workshop.
** These elements are optional and may be omitted if not relevant in a particular course.

Module 30: MEASURES OF RISK IN CASE-CONTROL STUDIES

OBJECTIVES

At the end of this session you should be able to:

1. **Define** incidence, risk and relative risk
2. **Estimate** relative risk from a case-control study using the appropriate measure:
   - Odds ratio for unpaired observations
   - McNemar's estimate of relative risk for paired observations

I. **Introduction**

II. **Incidence, Risk, and Relative Risk**

III. **Estimating Relative Risk in a Case Control Study**

I. INTRODUCTION

In a case-control study the investigator compares a group of cases, among which the problem that he wishes to investigate is present (e.g., a disease), with a group of controls, among which the problem is absent. The researcher makes this comparison to determine what factors may have contributed to the problem.

Before we examine the appropriate analysis, we will discuss a few important concepts.

II. INCIDENCE, RISK, AND RELATIVE RISK

There is no general agreement on the definitions of incidence and risk, but in this module we will use the following definitions:

> INCIDENCE is the total number of *new* cases of a defined condition (for example a disease) that occur during a specified period of time in a defined population.

Example:

The total number of new tuberculosis cases in District A in 1987 was 273. We may say that the **incidence** of tuberculosis in District A in 1987 was 273.

> INCIDENCE RATE is the total number of new cases of a defined condition that occur during a specified period of time divided by the "population at risk."

Example:

District A has a population of 200,000. The **incidence rate** of tuberculosis in 1987 in District A was, therefore, 273/200,000 per year or 137/100,000 per year.

> RISK is the same as incidence rate.

Example:

The **risk** of getting tuberculosis in District A in 1987 is 137/100,000 per year.

The risk may not be the same for various subgroups in the population. Whereas the risk of getting tuberculosis for farmers might be 100/100,000 per year, it may be 200/100,000 per year for mine workers. In this example mine workers are twice as likely to get tuberculosis.

It may be concluded that being a mine worker is a **risk factor** for contracting tuberculosis and carries a **relative risk** of 2.

A RISK FACTOR is any factor whose presence is associated with an increased risk of a disease or condition.

It is important to note that a risk factor does **not** imply that there is a **causal relationship** between the factor and the condition.

When determining relative risk we have to consider two subgroups in the study population: a subgroup in which the risk factor is present and one in which the risk factor is absent.

RELATIVE RISK is the risk of getting the disease in the group with the risk factor divided by the risk of getting the disease in the group without the risk factor.

Note that the higher a relative risk is, the more likely it becomes that the risk factor is causal and not due to confounding variables.

III. ESTIMATING RELATIVE RISK IN A CASE-CONTROL STUDY

As mentioned earlier, case-control studies are conducted to identify risk factors for diseases or conditions. When analyzing the results of a case-control study it is necessary to construct analytic cross-tabulations in which the cases and controls are placed in columns while the different variables which are considered as possible risk factors are placed in rows. (See **Module 23**.)

If a difference is observed between cases and controls with regard to a specific variable (risk factor), a chi-square test can be performed to determine if this difference is statistically significant. (See **Module 27**.)

However, when this test is performed, the strength of the relationship between the two variables (having or not having the risk factor and having or not having the disease) is not measured. Therefore, one needs a measure of association. The RELATIVE RISK is such a measure of association.

This analysis is helpful to us in solving practical problems. If you know of a risk factor and among whom it is present and you also know its relative risk, you can estimate to what extent the incidence of the disease can be reduced by preventive measures (assuming that the risk factor is causal).

For example:

If you know that smokers are 10 times more likely to develop lung cancer than non-smokers, you may assume that a health education campaign that brings down the percentage of smokers in an adult population from 40% to 35% would lead to a dramatic decrease in the incidence of lung cancer.

In case-control studies the relative risk often cannot be calculated because the incidence of the disease in the total population, from which the sample was drawn, is not measured. However it is possible to **estimate** the relative risk.

We will now discuss procedures for obtaining estimates of relative risk for two different situations: for unpaired observations and for paired observations.

Estimating Relative Risk - Unpaired Observations

For unpaired observations, the relative risk can be estimated by the ODDS RATIO. However, this is true only if two conditions are met:

- The disease has a low incidence in the total population, in both the at-risk group and the not-at-risk group. Under 0.05 will usually suffice.
- The control group is representative of the total population.

Example 1:

In a case-control study on smoking as a risk factor for lung cancer the following data were obtained:

Table 30.1. Smokers and nonsmokers among cases and controls.

	Lung cancer cases	Healthy controls
Smokers (+)	350	184
Nonsmokers (-)	45	216
Total	395	400

We determine from this table that more lung cancer cases smoke than healthy controls. This cannot be explained by sampling variation ($\chi^2 = 164$; d.f.= 1; $p < 0.001$).

The relative risk for getting lung cancer that is associated with smoking cigarettes is estimated by calculating the ODDS RATIO:

$$\text{Odds ratio} = \frac{350 \times 216}{184 \times 45} = 9.1$$

This means that those who smoked in our study were 9.1 times as likely to develop lung cancer as those who did not smoke.

The steps for performing this general type of analysis are presented below:

Step 1. Prepare a table:

Risk factor	Cases	Controls	Total
Present (+)	a	b	e
Absent (-)	c	d	f
Total	g	h	n

Step 2. Examine whether a difference exists between cases and controls with respect to the variable that is suspected to be a risk factor. Perform a χ^2-test to determine whether the difference is statistically significant.

Step 3. Estimate the relative risk by calculating the odds ratio. The odds ratio is a ratio of two ratios. It is the ratio of cases who have the risk factor (a) to cases who do not have the risk factor (c), divided by the ratio of controls who have the risk factor (b) to controls who do not have the risk factor (d). In a formula:

$$\text{Odds ratio} = \frac{\text{odds of cases who have risk factor}}{\text{odds of controls who have risk factor}} = \frac{a/c}{b/d} = \frac{ad}{bc}$$

Notes:

- When applying the above formula for the odds ratio, make sure that the table from which you take the data is in the same format as the one presented above.
- Do not calculate the odds ratio if the χ^2-test shows that the difference between cases and controls is not statistically significant.

Estimating Relative Risk - Paired Observations

In some case-control studies a matched control is selected for each case. In the analysis, we have to take account of this. We will once again use the example that was presented in **Module 28.**

Example 2:

In a case-control study carried out to determine causes of a cholera outbreak in Italy the following table was prepared:

Table 30.2. Consumption of fish or seafood by cholera patient/control pairs in 5 days preceding illness.

	Cholera cases		
Healthy controls	Ate seafood	Did not eat seafood	Total
Ate seafood	12	3	15
Did not eat seafood	30	31	61
Total	42	34	76

Source: Baine, W.B., Mazzotti, M, Greco, D, Izzo, E., Zampieri, A., Angioni, G., Di Gioia, M., Gangarosa, E.J., Pocchiari, S. 1974. Epidemiology of cholera in Italy in 1973. Lancet, ii (Dec.), 1370-1374.

For each cholera case (bacteriologically confirmed), a control was sought who was of the same sex, was of the same age decade, and was living in the same neighbourhood.

We find that, in 30 pairs, the cases ate seafood while the controls did not, whereas in only 3 pairs controls ate seafood while cases did not. As we saw already in **Module 28**, this association is not due to sampling variation (McNemar's $\chi^2 = 20.5$; $p < 0.001$).

In this case, the relative risk is estimated as:

$$\text{Relative risk} = \frac{30}{3} = 10$$

In other words, those who ate seafood were 10 times more likely to get cholera than those who did not eat seafood.

The steps for performing this general type of analysis are described below:

Step 1. Prepare a table:

	Cases	
Controls	Risk factor +	Risk factor -
Risk factor +	q	r
Risk factor -	s	t

Step 2. Examine whether an association exists between the risk factor and the disease or condition. This is done by comparing r and s in the table. Perform a McNemar's χ^2-test to determine whether the association is statistically significant (i.e., is not due to sampling variation).

Step 3. Estimate the relative risk by using the following formula:

$$\text{Relative risk} = \frac{s}{r}$$

GROUP WORK

If your study is a case-control study, estimate the relative risk by calculating the odds ratio (if observations are unpaired) or the ratio of the numbers of discordant pairs (if observations are paired) for different risk factors.

Remember that you can only do this for those risk factors where the difference between cases and controls has been found to be statistically significant.

Annex 30.1. Estimating Relative Risk Taking into Account the Effect of a Confounding Variable

A case-control study is carried out in a health centre to determine whether breast-feeding protects children aged 12-23 months against malnutrition.

The following results are obtained:

Table 30.3 Results of case-control study of whether breast-feeding protects children aged 12-23 months against malnutrition.

	Cases (malnourished)	Controls (well nourished)	Total
Not breast-fed (-)	100	76	176
Breast-fed (+)	100	124	224
Total	200	200	400

$\chi^2 = 5.36$; 1 d.f.; $p < 0.05$; Relative risk $= \frac{100 \times 124}{100 \times 76} = 1.6$

It is suspected that age is a confounding variable because:

- children aged 12-17 months are more often malnourished than those aged 18-23 months, and
- more children aged 12-17 months are breast-fed than those aged 18-23 months.

Ideally, age would have been taken into account in the design of the study, but this unfortunately has not been done.

The question now is: **How do we analyze these data?**

The answer is obtained by performing three steps:

Step 1. We split up the table into two parts:

Table 30.4. Results for children aged 12-17 months.

	Cases	Controls	Total
Not breast-fed (-)	50 (42%)	22 (28%)	72
Breast-fed (+)	70	58	128
Total	120 (100%)	80 (100%)	200

$\chi^2 = 3.58$; 1 d.f.; $p > 0.05$; Relative risk = 1.9

Table 30.5. Results for children aged 18-23 months.

	Cases	Controls	Total
Not breast-fed (-)	50 (63%)	54 (45%)	104
Breast-fed (+)	30	66	96
Total	80 (100%)	120 (100%)	200

$\chi^2 = 5.2$; 1 d.f.; $p < 0.05$; Relative risk = 2.0

You will note that for each of the subgroups the relative risk of not breast-feeding is higher than that obtained for the group overall (approximately 2). We would like to apply an overall significance test and estimate the overall relative risk, so we proceed with the following steps.

Step 2. We apply an overall significance test: the Mantel-Haenszel χ^2. (See **Annex 27.3.**)

	O_a	E_a (eg/n)	V_a (egfh/n^2(n - 1))
12-17 months	50	43.2	11.1
18-23 months	50	41.6	12.0
Total	100	84.8	23.1

$$\chi^2 = \frac{(O_a - E_a - 0.5)^2}{V_a} = \frac{(100 - 84.8 - 0.5)^2}{23.1} = 9.35;\ 1\ \text{d.f.};\ p < 0.01$$

Step 3. The overall estimate of relative risk is calculated as follows:

	ad/n	bc/n
12-17 months	14.5	7.7
18-23 months	16.5	8.1
Total	31.0	15.8

$$\text{Relative risk} = \frac{ad/n}{bc/n} = \frac{31.0}{15.8} = 2.0$$

Trainer's Notes

Module 30: MEASURES OF RISK IN CASE-CONTROL STUDIES

Timing and Teaching Methods

1 hour	Introduction and discussion
2 hours	Group work

Introduction and Discussion

- To make the presentation easier to follow, do not merely use the terms cases and controls, but talk instead about ill persons and healthy persons or cholera cases and healthy controls.
- If any of the groups has a case-control study, make use of an example from the study to illustrating how to calculate an odds ratio and its interpretation.
- **Annex 30.1** should not be presented. It can be used by the groups, if necessary.
- If none of the groups has paired observations and if the participants have little experience with statistics, the latter part of the section on estimating relative risk - paired observations can be left out of the presentation.

Health Systems Research Training Series

Volume 2, Part II: Data Analysis and Report Writing

Module 31:

REPORT WRITING

STEPS IN DATA ANALYSIS AND REPORT WRITING

Questions you must ask	Steps you will take*	Important elements of each step
What data have been collected for each research objective? Are data complete, accurate?	Prepare data for analysis	Review field experiences Inventory data for each objective/study population Sort data and check quality Check computer outputs
What do the data look like?	Describe variables	Frequency distributions Figures, means
How can the data be summarized for easy analysis?	Cross-tabulate quantitative data Summarize qualitative data	Cross-tabulate in relation to objectives Graphic displays, narratives
For quantitative data: does each research objective aim to describe, compare, or find associations? 1. How can the data be described?	Determine the type of statistical analysis required	Review objectives, study type, and variables Statistical description of variables Choosing significance tests
2. How can differences between groups be determined?	Analyze paired and unpaired observations	**Student's t-test **Paired t-test **Chi-square test **McNemar's chi-square test
3. How can the associations between variables be determined?	Measure associations between variables	**Scatter diagram **Regression line and correlation coefficient **Relative risk, odds ratio
How should the report be written?	Write the report and formulate recommendations	Prepare outline for report Draft and redraft Summarize findings Summarize conclusions for each objective Formulate recommendations Prepare abstract
How should the findings and recommendations be presented and disseminated?	Present a summary and draft a plan for implementation of recommendations	Discuss summaries with different target groups Discuss plan for implementation

* These steps need not be in the sequence in this diagram. The sequence may be adjusted according to the needs of the workshop.
** These elements are optional and may be omitted if not relevant in a particular course.

Module 31: REPORT WRITING

OBJECTIVES

After this session, you should be able to:

1. List the main components of a research report.
2. Make an outline of your research report.
3. Write drafts of your report in stages.
4. Check the final draft for completeness, possible overlaps, and for clarity and smoothness of style.
5. Draft recommendations for action based on your research.

I. **Steps in Preparing the Report: Preliminary Considerations**

II. **Writing the Report**

I. STEPS IN PREPARING A REPORT: PRELIMINARY CONSIDERATIONS

The Audience

The purpose of a research report is to convey information to the reader. Therefore, it is important to begin by clarifying in your mind:

WHO is the reader?
WHY does he or she want to read the research report?

In health systems research, it is particularly important to remember the needs of the audience because the audience is not only the research community, but also health managers and community leaders. Many research papers that are meant for scientific people are not suitable for managers and lay people. Therefore, special attention should be devoted to preparing reports that are simply worded and are explicit regarding findings.

Furthermore, it is important to present not only the scientific findings, but also specific recommendations that take into consideration the local characteristics of the health system, constraints, feasibility, and usefulness of the proposed solutions. The community and the manager are more interested in learning "what to do about a problem" than in being told "there is a problem."

> Reports should meet the **NEEDS OF THE AUDIENCE** of community leaders, health managers, and researchers.

How the Reader Reads a Research Report

Recognizing the "reading strategies" of people who read research reports will help you write a good report. The research was done to provide new information. Therefore, this should be the highlight and focus of the report. This "new information" should be summarized as the **conclusions** of the study. Most readers will begin by reading the conclusions. If this section is interesting, useful, and attractively presented, the reader will look at the other sections. The other sections of the report are intended to support the conclusions by helping the reader clarify two basic questions in his or her mind:

- **How will this "new information" help improve the health of the community?** (i.e., What is the problem and the health system in which the problem occurs and how will this information help solve or reduce the problem?)

- **Can I "believe" these findings?** (i.e., Are the findings valid and reliable?) The research design, sampling, methods of data collection, and the data analysis will substantiate the validity and reliability.

Note that a report that highlights the methodology sections rather than the conclusions might interest a researcher audience, but will not interest the manager audience.

Completing the Data Analysis

Before you begin the outline and first draft of your report, you need to review your analysis of the data asking several of the following questions:

- **Are conclusions appropriate to the specific objectives? Are they comprehensive?**

  The earlier steps in data analysis should have produced:

  - one or more conclusions stated as simple sentences; and
  - one or more analytic tables together with the relevant descriptive statistics or statistical tests to support the conclusions.

  Review these conclusions and check whether:

  - every specific objective has been dealt with;
  - all aspects of each objective have been dealt with; and
  - the conclusions are relevant and appropriate to the objectives.

- **Are further analytic tables needed?**

  If the conclusions are not comprehensive, prepare further dummy analytic tables and analyze the data as described in **Modules 22-30.**

- **Have all qualitative data been used to support and specify conclusions drawn from tables?**

Once you have completed this review, you need to complete a couple of additional tasks:

- **State the final conclusions in relation to each objective.**

  During earlier stages of analysis, every analytic table would have had a conclusion. These conclusions should now be reviewed, combined whenever possible, and stated in such a way that the main findings of the study are easily identifiable by a reader who is "scanning" the report. Very often the most important numerical information (%, means etc.) can be included in these statements.

- **Select supportive tables to appear in the text of the report.**

  The number of tables in the body of the report should be very limited. A table should be included only if it illustrates an important conclusion or provides evidence to support it. When possible, combine information from several analytic tables into one or more and present a summary table in the body of the report. (If necessary, more detailed tables can be placed in annexes.) The title of each table should tell the reader in as few words as possible exactly what the table contains. Column and row headings should be brief, but self-explanatory.

Compile the conclusions and tables relating to each specific objective. You are now ready to draft the report.

II. WRITING THE REPORT

The aim of the report is to tell the reader the facts in a simple, logical, sequential fashion. Avoid confusion and distracting the reader.

In writing the report, it is important to consider:

- the CONTENT,
- the STYLE of writing,
- the LAYOUT of the report,
- FIRST DRAFT,
- SECOND DRAFT,
- finalizing the report.

Each of these aspects of report preparation will be discussed in turn.

Content: Main Components of a Research Report

The research report should contain the following components:

Title or cover page,
Summary of findings and recommendations,
Acknowledgments (optional),
Table of contents,
List of tables, figures (optional),
List of abbreviations (optional),
1. INTRODUCTION,
2. OBJECTIVES,
3. METHODOLOGY,
4. FINDINGS AND CONCLUSIONS,
5. DISCUSSION,
6. RECOMMENDATIONS,

References,
Annexes (data collection tools, tables).

The findings and conclusions, discussion of findings, and recommendations will form the most substantial part of your report, which has to be written from scratch. For the introduction you can rely to a large extent on your research proposal, although you may summarize, revise, and sometimes expand certain sections.

We, therefore, strongly advise that you **start with the findings and conclusions.** Nevertheless we will briefly elaborate on each component in the sequence in which they will finally appear in your report.

- **Cover Page**

  The cover page should contain the title, the names of the authors with their titles and positions, the institution that publishes the report, and the month and year of publication. The institution that publishes the report will most likely be the one that administered the project, for example, the Research Unit of the Ministry of Health or a research institute.

- **Summary**

  The summary can only be written after the first or even the second draft of the report has been completed. It should contain:

  a very brief description of the problem (WHAT),

- the main objectives (WHY),
- the place of study (WHERE),
- the type of study and methods used (HOW),
- the main findings and conclusions, followed by
- the major, or all, recommendations.

The summary will be the first (and for busy health decision-makers most likely the only) part of your study that will be read. Therefore, its writing demands thorough reflection and is time consuming. Several drafts may have to be made, each discussed by the research team as a whole.

As you will have collaborated with various groups during the drafting and implementation of your research proposal, you may consider writing different summaries for each of these groups. For example, you may prepare different summaries for policymakers and health managers, for health staff of lower levels, for community members or the public at large (newspaper, TV), and for professionals (articles in scientific journals). (See **Module 32**.)

- **Acknowledgments**

  You may wish to thank those who supported you technically or financially in the design and implementation of your study. Also your employer, who has allowed you to invest time in the study, and the respondents may be acknowledged. Acknowledgments are usually placed right after the cover page or at the end of the report, before the references.

- **Table of Contents**

  A table of contents is essential, as it gives the reader a quick overview of the major sections of your report, and page references, if he wishes to go through the report in a different order or skip certain sections.

- **List of Tables, Figures (optional)**

  If you have many tables or figures it is helpful to list these also, in a "table of contents" type format with page numbers.

- **List of Abbreviations (optional)**

  If there are many abbreviations or acronyms in the report, these could be listed in addition.

The latter three sections should be prepared last, as you have to include the page numbers of all chapters and subsections in the table of contents, and be sure there are no mistakes in the final numbering of figures and tables.

1. **INTRODUCTION**

   The introduction is a relatively easy part of the report which may be written after a first draft of the findings has been made. It should certainly contain some background data about the country, the health status of the population, and health service data related to the problem that has been studied. You may slightly revise or make additions to the corresponding section in your research proposal and use it here.

Then the statement of the problem should follow, again revised from your research proposal with comments or additional data based on your research experience added, if useful. It should contain a paragraph on what you hope to achieve with the results of the study.

A brief review of the literature pertaining to your topic of study should then be given. (Consult **Module 5** and your research proposal.) This section should include relevant points to help the reader:

- understand the problem providing a review of available information on it, and
- understand methods of investigating or resolving the problem.

NOTE: This section should NOT be a summary of all the papers and books on the topic. Be selective, remembering that this section serves to lend support for your study, not to display your ability to read literature.

2. OBJECTIVES

The general and specific objectives should be included. If necessary, you can adjust them slightly for style and sequence. However, you should not change their basic nature. If you have not been able to meet some of the objectives, this should be stated in the methodology section and in the discussion of the findings.

3. METHODOLOGY

The methodology you followed for the collection of your data should be described in detail. It should include :

- the study type,
- the variables on which data was collected,
- the population from which the sample was selected,
- the size of the sample and method of sampling,
- the data collection techniques:
  - sources of data (cards, households, clinic registers, etc.),
  - how the data was collected and by whom,
  - procedures for data analysis, including statistical tests (if applicable).

If you have deviated from the original study design presented in your research proposal, you have to explain to what extent and why. The consequences of this deviation for meeting certain objectives of your study should be indicated. If the quality of some of the data is weak, resulting in possible biases in a certain direction, this should be described.

4. FINDINGS AND CONCLUSIONS

The systematic presentation of your findings and conclusions in relation to the research objectives is the crucial part of your report.

A description of the findings may be complemented by a limited number of tables or graphs that summarize the findings. The text will become more lively if you illustrate some of the findings with examples using the respondents' own words, or with observations and case-studies that you recorded during the fieldwork.

5. DISCUSSION

The findings can be discussed by objective or by cluster of related variables. The discussion should also mention findings from other related studies that support or contradict your own. It is important, as well, to present and discuss the limitations of the study. In the discussion of findings some general conclusions may be included as well.

Note:

The text and annexes should include sufficient details for professionals to enable them to follow how you substantiate your findings and conclusions. The report should be so self-explanatory that it should be possible to repeat the study, if desired.

6. RECOMMENDATIONS

The recommendations should follow logically from the discussion of the findings. They may be summarized according to the groups toward which they are directed, for example:

- policymakers,
- health and health-related managers at district or lower level,
- health and health-related staff who could implement the activities,
- potential clients, and
- the community at large.

Remember that action-oriented groups are most interested in this section.

In making recommendations, use not only the findings of your study, but also supportive information from other sources and available information on other related factors. The recommendations should be discussed with all concerned before they are finalized.

If your recommendations are short, you might include them all in your summary of findings and recommendations and omit them as a separate section.

- **References**

The references in your text can be numbered in the sequence in which they appear, then listed in this order in the reference section. Another possibility is to list the author's names in the text followed by the date of the publication in brackets, for example (Shan 1990). In the list of references, the publications are then arranged in alphabetical order by the principal author's last name (see Module 5).

You can choose either method, but if you wish to publish an article you must follow the method used in the journal to which you wish to submit your article.

- **Annexes or Appendices**

The annexes should contain any additional information needed to enable professionals to follow your research procedures and data analysis.

Information that would be useful to special categories of readers but is not of interest to the average reader could be included in annexes, as well.

Examples of information in annexes are:

- tables referred to in the text but omitted to keep the report short;
- lists of criteria, definitions, and flow-charts;
- lists of hospitals, districts, villages, etc., that have participated; and
- all data collection tools.

Style of Writing

Remember that your reader:

- Is short of time,
- Has many other urgent matters demanding his or her interest and attention, and
- Is probably not knowledgable concerning "research jargon."

Therefore the rules are:

- **Simplify**. Keep to the essentials.
- **Justify**. Make no statement that is not based on facts.
- **Quantify**. Avoid "large," "small"; instead, say "almost 75%," "one in three," etc.
- **Be precise and specific**.
- **Inform, not impress.** Avoid exaggeration.
- **Use short sentences**.
- **Use adverbs and adjectives sparingly**. Be consistent in the use of tenses (past, present tense). Avoid the passive voice, if possible.
- **Aim to be clear, logical, and systematic in your presentation**.

Layout of the Report

A good physical layout is important as it will help your report:

- Make a good initial impression,
- Encourage the reader, and
- Give an idea of how the material has been organized so the reader can make a quick determination of what he or she will read first.

Particular attention should be paid to make sure there is:

- An attractive layout for the title page; a clear table of contents,

- Consistency in margins and spacing,
- Consistency in headings and subheadings, (e.g., **capitals, underlined**, for headings of chapters; **capitals** for headings of major sections; **lower case, underlined**, for headings of subsections, etc.),
- Good quality typing and photocopying. Correct drafts carefully. (For more detailed information, see Keithly and Schreiner 1971).
- Numbering of figures and tables, provision of clear titles for them, and labels for columns and rows, etc.,
- Accuracy and consistency in quotations and references.

Preparing the First Draft of the Report

Prepare a written OUTLINE of the report. An outline will help to organize your thoughts and is an essential step in producing a logical, sequential report.

An outline should contain:

- Headings of the main sections of the report,
- Headings of subsections,
- Points to be made in each section, and
- A list of tables and figures (if relevant) to illustrate each section.

The outline for the chapter on findings and conclusions is the most difficult. A discussion with your team members concerning the main findings and conclusions of your data in relation to objectives and variables should help you structure your findings in a logical and coherent way.

- The first section under findings and conclusions is usually a description of the sample, for example in terms of location, age, sex, and other relevant background variables.
- Then, depending on the study design, you may provide more information on the problem or dependent variable(s) of your study.
- Next an analysis of the different independent variables in relation to the problem may follow.

You might start by listing headings and subheadings, with ample space between them so that you can scribble key words related to what you intend to write under each heading. It is advisable to number sections and subsections as you list them.

For example, in a study on malnutrition, the chapter "Findings and Conclusions" may look like this:

CHAPTER 4: FINDINGS AND CONCLUSIONS

4.1 DESCRIPTION OF THE SAMPLE

4.2 EXTENT AND SEASONAL VARIATION OF MALNUTRITION IN DISTRICT X

4.3 POSSIBLE CAUSES OF MALNUTRITION

4.3.1 Limited availability of food

4.3.2 Non-optimal utilization of food

4.3.3 High prevalence of communicable diseases

4.3.4 Limited access to MCH and curative services

4.3.5 Conclusions

This system of numbering is flexible and can be extended according to need with further headings or subheadings. It allows you to keep an overview of the process when different group members work on different parts at the same time. If your findings are very elaborate so that you get sub-sub-subheadings with 4 or 5 numbers, you might decide to split up the findings into several chapters. (In addition, you may consider leaving off some of the numbering on subsections, if it's clear under what major heading they belong. However, keep all the numbering until the final draft, as it helps you keep your report in order when various members of the group are working on different sections.)

TABLES and FIGURES in the text need numbers and clear titles. It is advisable to first use the number of the section to which the table belongs. In the last draft you may decide to number tables and figures in sequence.

Include only those tables and figures that present main findings and need more elaborate discussion in the text. Others may be put in annexes, or, if they don't reveal interesting points, be omitted.

Note that it is unnecessary to describe in detail a table that you include in the report. Only present the main conclusions.

The first draft is never final. Therefore you might concentrate primarily on content rather than on style. Nevertheless, it is advisable to structure the text straight from the beginning in paragraphs, and to attempt to phrase each sentence clearly and precisely.

Notes:

Never start writing without an outline. Make sure that all sections written carry headings and numbers consistent with the outline before they go for typing. Have the outline visible on the wall so everyone will be immediately aware of any additions or changes.

Type the first draft double-spaced with large margins so that you can easily make comments and corrections in the text.

Have several copies made of the first draft, so you will have one or more copies to work on and one copy on which to insert the final revisions you will hand in for retyping.

Preparing the Second Draft

When a first draft of the findings and conclusions has been completed, all working-group members and facilitators should read it critically and make comments.

The following questions should be kept in mind when reading the draft:

- Have all important findings been included?
- Do the conclusions follow logically from the findings? If some of the findings contradict each other, has this been discussed and possibly explained? Have weaknesses in the methodology, if any, been revealed?
- Are there any overlaps in the draft that have to be removed?
- Is it possible to condense the content? In general a text gains by shortening. Some parts less relevant for action may be included in annexes. Check if descriptive paragraphs may be shortened and introduced by a concluding sentence.
- Do data in the text agree with data in the tables? Are all tables consistent (with the same number of informants per category), are they numbered in sequence, and do they have clear titles?
- Is the sequence of paragraphs and subsections logical and coherent? Is there a smooth connection between successive paragraphs and sections? Is the phrasing of findings and conclusions precise and clear?

The original authors of each section may prepare a second draft, taking into consideration all comments that have been made. However, you might consider the appointment of two editors amongst yourselves, to draft the complete version.

In the meantime, other group members may (re)write the introductory sections. The INTRODUCTION, OBJECTIVES, and METHODOLOGY sections of your original research proposal may be used, in many cases, after being reviewed and adjusted (see page 8).

Now a first draft of the summary can be written (see page 7).

Finalizing the Report

It is advisable to have one of the other groups and facilitators read the second draft and judge it on the points mentioned in the previous section. Then a final version of the report should be prepared. This time you should give extra care to the presentation: structure, style, and consistency of spelling.

Use verb tenses consistently. Descriptions of the field situation may be stated in the past tense (e.g., "Five households owned less than one acre of land.") Conclusions on data are usually in the present tense (e.g., "Food taboos hardly have any impact on the nutritional status of young children. Those species of fish and meat that are forbidden for certain clans rarely appear in the daily diet.")

For a final check on readability, you might skim through the pages and read the first sentences of each paragraph. If this gives you a clear impression of the organization and results of your study, you may conclude that you did the best you could.

GROUP WORK

1. **Make an outline for your report** on a flipchart, after reviewing your objectives, your sources of information and the outcomes of your data analysis. Take extra care that your chapter of findings has logical sections and subsections. Number proposed sections and subsections. Stick the outline to the wall in a visible place. Leave sufficient space between the lines for additions (more subdivisions, for example) and for changes.

2. **Start writing, beginning with the chapters on findings and conclusions**. Decide with your facilitator whether you will interpret the data presenting results by variable, by objective, or by study population. If you are unsure in the beginning which method of organizing the presentation will work best, record your findings and interpretations anyway. In the second draft you can decide how to reorganize and shorten the presentation. **Divide writing tasks** among subgroups of one or two persons.

3. **Develop the chapters on introduction** (background, statement and literature review), objectives, and methodology, adapting what you prepared for the proposal.

4. **Finally, develop the summary** following the outline given earlier in this session. Take at least half a day for this, working systematically.

5. **Keep track of progress** in writing and typing, making notes on the flipchart which has the outline of your report.

6. **Go over the first draft with the group as a whole** checking it for gaps, overlaps, etc., before the second draft is prepared. Have a facilitator from another group read the whole draft report before it goes for final typing.

References

Huth, E.J. 1982. How to write and publish papers in the medical sciences (2nd ed.). Williams & Wilkins, Baltimore, MD, USA.

Keithly, Schreiner. 1971. A manual of style for the preparation of papers and reports. South Western Publishing Co.

Osborn, M.L. 1971. The preparation of a scientific paper. British Journal of Medical Education, 5, 248-252.

Paton, A. 1979. Write a paper. In How to do it: 1. Articles published in the British Medical Journal. British Medical Association, London, UK. pp. 207-212.

Trainer's Notes

Module 31: REPORT WRITING

Timing and Teaching Methods	
1 hour	Introduction and discussion
Several days	Group work

Introduction and Discussion

- Put the outline for research reports on an overhead sheet and discuss it point by point. Stress that the findings and conclusions, the discussion and the recommendations will have priority.

- Take an example from one of the groups when presenting a possible outline for the chapter on findings and conclusions with headings and subheadings.

  Explain the system of numbering, making sure that you are consistent in the layout of headings and subheadings so that you can use the example to illustrate appropriate layout later on.

- Ask the participants to suggest the criteria they would use to judge their first draft, before you give guidelines.

- Use examples from the research proposals prepared by various groups when discussing how the statement of the problem, objectives, and methodology should be adapted for the final report.

  Pay attention to the need for changing the future tense used in the proposal into the present and past tense, if you suspect that some groups may overlook this aspect.

Group Work

- Make sure that all of the groups first make an outline for their reports, using the outline presented in the module as a starting point. Ask the groups to hang up the outlines so everyone in their teams can see them.

- The sections on findings and conclusions, discussion, and recommendations will take the most time. Some groups may find that the presentation of these sections would work best ordered somewhat differently. (For example, it may be more logical for certain projects, to have sections entitled: 4. Findings, 5. Discussion and Conclusions, and 6. Recommendations.) Let them know that they can use whatever outline and system for presentation of this material is most appropriate for their own reports. Discuss with the groups how this part of their presentation can be structured most logically.

- Writing should start with the findings and conclusions. Only when a reasonable draft is ready they should (re)write the introductory chapters.

- Make sure that all group members have some writing tasks, for example, by letting them write in pairs. If certain participants have never written, they might need intensive support. You can let them write several paragraphs and then rewrite the page with them, to provide them with an example.

- If groups have no writing experience, they will need explicit guidance concerning what points they should check on when they review their first draft, including the basic layout of the report.

- It is advisable to discuss ideas for possible recommendations during the write up of the findings and conclusions. These ideas should be recorded immediately (preferably on a flipchart) so they can be used when phrasing recommendations.

- All facilitators should comment on the summary section prepared by each group. Each facilitator should also read and comment on at least one complete draft report from another group before it goes for final typing.

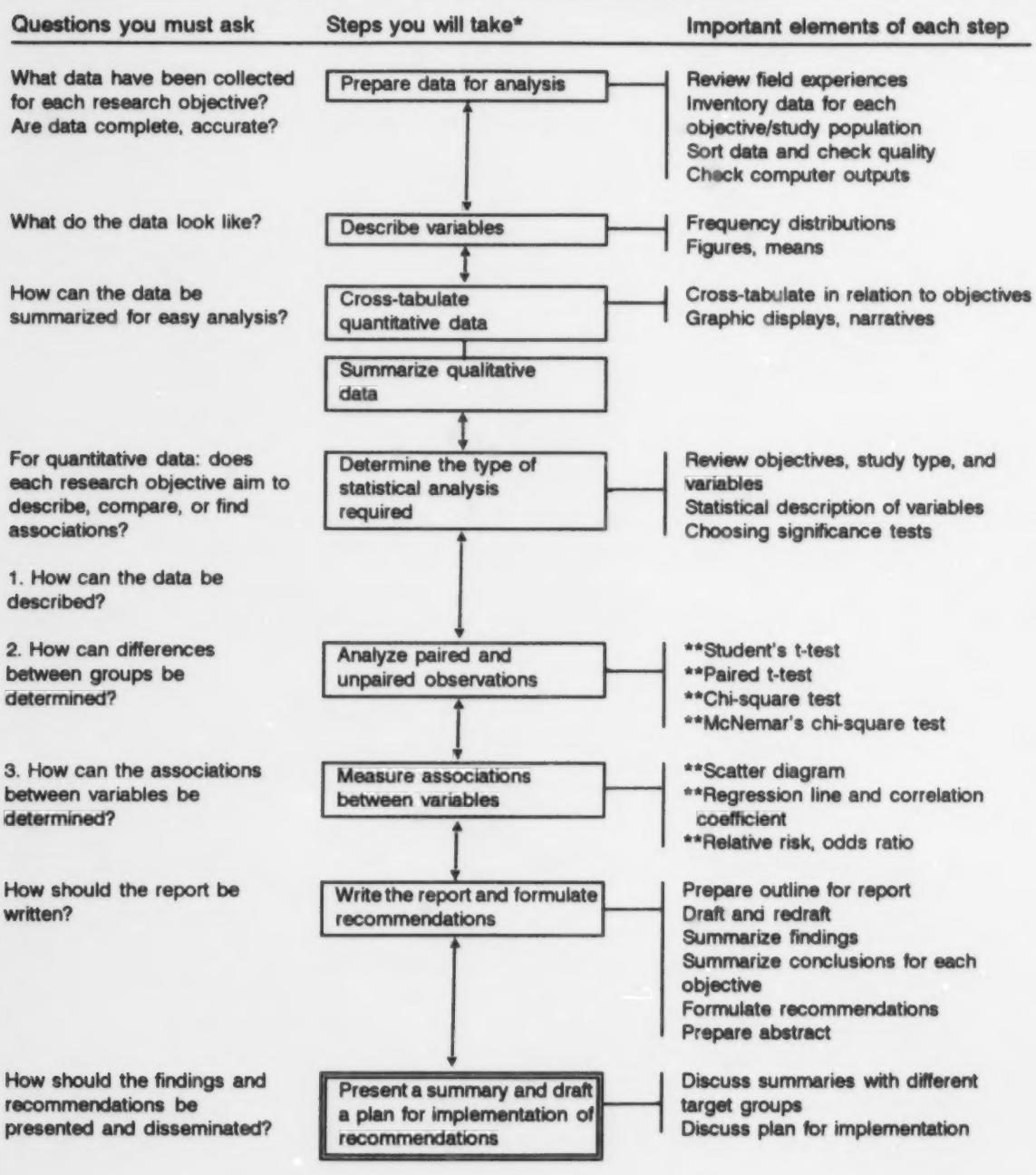
Health Systems Research Training Series

Volume 2, Part II: Data Analysis and Report Writing

Module 32:

PROMOTING THE UTILIZATION OF FINDINGS

STEPS IN DATA ANALYSIS AND REPORT WRITING


Questions you must ask
Steps you will take*
Important elements of each step
What data have been collected for each research objective? Are data complete, accurate?
Prepare data for analysis
Review field experiences
Inventory data for each objective/study population
Sort data and check quality
Check computer outputs
What do the data look like?
Describe variables
Frequency distributions
Figures, means
How can the data be summarized for easy analysis?
Cross-tabulate quantitative data
Cross-tabulate in relation to objectives
Graphic displays, narratives
Summarize qualitative data
For quantitative data: does each research objective aim to describe, compare, or find associations?
Determine the type of statistical analysis required
Review objectives, study type, and variables
Statistical description of variables
Choosing significance tests
1. How can the data be described?
2. How can differences between groups be determined?
Analyze paired and unpaired observations
**Student's t-test
**Paired t-test
**Chi-square test
**McNemar's chi-square test
3. How can the associations between variables be determined?
Measure associations between variables
**Scatter diagram
**Regression line and correlation coefficient
**Relative risk, odds ratio
How should the report be written?
Write the report and formulate recommendations
Prepare outline for report
Draft and redraft
Summarize findings
Summarize conclusions for each objective
Formulate recommendations
Prepare abstract
How should the findings and recommendations be presented and disseminated?
Present a summary and draft a plan for implementation of recommendations
Discuss summaries with different target groups
Discuss plan for implementation

* These steps need not be in the sequence in this diagram. The sequence may be adjusted according to the needs of the workshop.

** These elements are optional and may be omitted if not relevant in a particular course.

Module 32: PROMOTING THE UTILIZATION OF FINDINGS

OBJECTIVES

After this session you should be able to:

1. **Present** a summary of your findings and recommendations to health managers and other interested parties.
2. **Draft** a plan of action to promote implementation of the recommendations.

I. **Introduction**

II. **Presenting the research results to health managers**

III. **Presenting the major findings and recommendations to other interested parties**

IV. **Drafting a plan of action**

I. INTRODUCTION

There are several ways to promote the utilization of your research findings:

- Presenting the research results to health managers and community leaders,
- Presenting the research results to other interested parties,
- Drafting a plan of action to promote the implementation of the recommendations that resulted from your study.

A combination of these three options may be the most effective way to promote utilization of your results.

II. PRESENTING THE RESEARCH RESULTS TO HEALTH MANAGERS

The purpose of HSR is to provide information to managers that will facilitate decision-making. Therefore an extremely important step in the health systems research process is the presentation of the report to the appropriate managers so you can discuss with them the findings and recommendations. This should be a face-to-face presentation in which there is sufficient opportunity for discussion. Merely sending a copy of the report or research paper to the managers is usually not adequate.

To ensure that such presentations are useful and productive:

1. **Select the audience carefully.**

   Be sure that relevant managers and their key staff are invited. This may include, for example, clinical staff, supervisory staff, field staff, personnel from other sectors and interested agencies, and, if appropriate, community leaders.[3] It may be necessary to make more than one presentation and modify the style and content to suit each audience.

2. **Make sure that sufficient time is allocated for the presentation and for discussion.**

   A minimum would be half an hour for each. The discussion, however, could easily last longer if it covers the possibilities for implementing the recommendations.

3. **Arrange your presentation.**

   Your presentation should consist of:

   - A brief introduction, including the statement of the problem, objectives of the study, sample and data collection tools used.
   - The major findings, listed in a logical sequence (for example, starting with a description of the problem, followed by the major variables which influence it).
   - The recommendations, roughly following the same sequence.

---

[3] A separate presentation may be made to the community as a whole, or to patients that have participated in your study. (See **Section III.**) If community leaders are important decision-makers who may use your results and should participate in the dialogue with health managers concerning how the recommendations should be implemented, they should be invited to this first presentation.

Preferably separate recommendations should be prepared for policymakers, for health managers, for health staff, and for community members. This makes it easy to talk with each group.

Remember that your audience basically wants to know: "How can we solve this problem?" Therefore:

- Avoid technical jargon.
- Do NOT overload the audience with statistical data. Present relevant findings that require remedial action. (However, you might present one or two tables to support your main conclusions on overhead sheets, or illustrate them with an interesting observation.)
- Be very specific in your recommendations concerning the actions required to solve the problem. If your research indicates that there are several viable options, describe the alternatives and their potential advantages and disadvantages. (If you are presenting your study to those who may be sensitive, you may appear to be "telling them what to do." Instead, you might consider presenting options, illustrating clearly what information your findings provide that will be helpful in choosing an alternative, then let the managers and community make the final choice.)

4. **Prepare appropriate visual aids**

   Have sufficient copies of your typed Summary of Findings and Recommendations for all who are present. If the presentation takes place **after** you have completed your report, have some copies of the full report available for those who are most concerned or interested.

   Prepare overhead sheets, slides, or flipcharts to highlight the most important points in your presentation (e.g., problem, main objectives, major findings, and recommendations that require action from those you are addressing).

5. **Be prepared to discuss the logic and feasibility of the recommendations.**

   Also be prepared to accept suggestions, criticism, and reactions and make appropriate modifications.

6. **Appoint two team members as recorders for the session.**

   Make sure that proper minutes of the discussion are taken, especially concerning the decisions and follow-up action that is agreed upon. These minutes should subsequently be circulated to all those who were present, as well as to key persons who were invited, but did not personally attend.

GROUP WORK, PART I

1. Prepare your presentation for a selected group. Allow sufficient time for developing the presentation — at least half a day for drafting the outline of what your will present and any notes you will need and then adapting your "Summary of Major Findings and Recommendations" for distribution, if necessary. Reserve another half a day for typing, corrections, making copies of the summary and any handouts and preparing overhead sheets or flipcharts.

2. The Principal Investigator may make the entire presentation, but it is also possible to assign various parts of the presentation to different group members.

III. PRESENTING THE MAIN FINDINGS AND RECOMMENDATIONS TO OTHER INTERESTED PARTIES (EXAMPLES)

Community or Patients

If your research pertains to a problem that affects community members or patients and if you have interviewed a number of them, it is recommended that you report your major findings and recommendations to them as well.

Your presentation may serve several purposes:

- to inform community leaders, patients, or the community at large about your findings;
- to check whether they agree with your conclusions concerning the nature, magnitude, and causes of the problem, based on their own experiences;
- to solicit additional information on questions that remain;
- to inform them briefly about recommendations for policymakers and managers.
- to concentrate on recommendations that concern them, to obtain their opinion on appropriateness and feasibility and elicit their support for any actions they themselves should take; and
- to solicit additional suggestions for action.

Other Researchers

Although action is the first aim of your research, its scientific importance should not be neglected. Researchers in community health and social science departments, for example, might be interested in the methodology you used and your research results. It is, therefore, important to invite them to an official presentation of your findings, or to communicate your findings during symposia or meetings organized by interested research institutions.

Also publication in a local or international research bulletin or journal should be considered.

The Mass Media

Communicating research findings to the mass media is a good strategy for promoting interest in and support for research. It is important to remember that it is information that has a human-interest angle that is attractive to the mass media. Therefore, the researcher can highlight such aspects. However, it is also important to avoid antagonizing health managers, health care providers, or the community in the process of releasing information to the mass media. This could jeopardize the use of the findings and future collaboration. One effective strategy is to involve the appropriate health managers and community leaders in releasing the results.

IV. DRAFTING A PLAN OF ACTION

When developing your research proposal, you discussed as a group what concrete results you expected to obtain from your study and how the utilization of these results could be promoted.

An important step has been taken by involving health managers as well as the target population in the formulation of the recommendations.

Now you should proceed to identify how you can assist in the implementation of these recommendations. Review the section of your research proposal in which you already made some suggestions concerning how your findings might be utilized and elaborate or adjust them.

1. **Discuss the following questions:**

   - Which recommendations can you implement yourself without further authorization or extra support?
   - How would you proceed with the implementation?
   - For which recommendations is the support of other authorities required? (Specify the authorities as well as the support needed.)
   - In what ways can you encourage this support?

2. **Draft a plan of action** for the coming year, or, if possible, for a longer period, indicating:

   - Activities that can take place without further authorization required, their timing, who will be responsible, and how they will be carried out;
   - Suggestions for further actions and steps to be taken to encourage their implementation.

3. If there are a number of specific actions that should be taken for which detailed planning by several parties is needed, you might consider **holding an "action planning workshop."** In this setting the various groups involved could work together to study the findings and recommendations in detail and develop an "action plan." The presentation for managers and community leaders could be expanded to include this active planning phase, or a separate workshop could be scheduled, after key decisions-makers have had time to review the results and their implications.

   The workshop could be anywhere from a couple hours to a couple days in length, depending on the size of the study and the nature of the actions that need to be planned. If a longer format is chosen, you might consider a short field visit to some of the research sites before the group begins the "action planning" phase of the workshop.

   When plans are drafted, be sure the working groups consider what activities and tasks will be completed, who will be responsible, when they will take place, and what resources are needed.

GROUP WORK, PART II

1. If applicable, plan a presentation for community members or patients and determine how you will obtain feedback concerning your research findings and recommendations.

2. Draft a plan for implementing your study results.

3. If the plan for implementing results should involve the participation of several individuals or organizations, consider whether an "action planning workshop" would be useful. If so, make preliminary plans for it.

N.B. While the final version of your report is being typed you may have time to carry out this group work. Otherwise it will have to be completed after the workshop is over. Try to do as much of this planning as you can while you are still at the workshop.

Trainer's Notes

Module 32: PROMOTING THE UTILIZATION OF FINDINGS

Timing and Teaching Methods

20 minutes Introduction and discussion

Introduction and Discussion

It is important to present this module before participants start preparing the summary of their findings and recommendations.

Note:

Stress that the written summary should be added to the report just after the acknowledgments.

Group Work

When presenting the research projects it is recommended that **participants distribute the summary including the full recommendations to all course participants and guests of honour**, to enhance their participation in the discussion. Details on the specific objectives of the study, the methodology, and the findings (for example, some crucial tables or graphs) can be presented using flip charts or overhead sheets.

Head Office
IDRC, PO Box 8500, Ottawa, Ontario, Canada K1G 3H9

Regional Office for Southeast and East Asia
IDRC, Tanglin PO Box 101, Singapore 9124, Republic of Singapore

Regional Office for South Asia
IDRC, 11 Jor Bagh, New Delhi 110003, India

Regional Office for Eastern and Southern Africa
IDRC, PO Box 62084, Nairobi, Kenya

Regional Office for Middle East and North Africa
IDRC, PO Box 14 Orman, Giza, Cairo, Egypt

Regional Office for West and Central Africa
IDRC, BP 11007, CD Annexe, Dakar, Senegal

Regional Office for Latin America and the Caribbean
IDRC, Casilla de Correos 6379, Montevideo, Uruguay

Please direct requests for information about IDRC and its activities to the IDRC office in your region.

ABOUT THE AUTHORS

Corlien M. Varkevisser, PhD, MPH, is a medical sociologist-anthropologist by profession who specialized in public health. As a staff member of the Royal Tropical Institute, Amsterdam, and former head of the Primary Health Care (PHC) Unit, she has gained extensive experience in health systems research and PHC management in sub-Saharan Africa. She is one of the coinitiators of the Joint HSR Project (WHO/Netherlands Ministry for Development Cooperation/Royal Tropical Institute) for southern Africa and has been based at the WHO subregional office in Harare as manager of the Joint HSR Project since its onset in April 1987.

Indra Pathmanathan, MMBS, MPH, is a physician specializing in public health who is currently working in the Ministry of Health, Malaysia. She was previously on the academic staff of the University of Malaya. As head of the HSR program in Malaysia since its inception, she has been responsible for developing and implementing several strategies for HSR that have been replicated in other countries. These included training programs in HSR and Quality Assurance for decision-makers in ministries, for physicians, and for others in district health teams, hospitals, and universities. She is a member of the Advisory Group on HSR, WHO-Geneva and serves on the editorial board of BRIDGE.

Ann Brownlee, MA, PhD, is a medical sociologist who specializes in HSR, planning and evaluation, and cross-cultural aspects of health care. She served as Research and Evaluation Coordinator for the Project for Strengthening Health Delivery Systems in West and Central Africa for a number of years, where she worked closely with WHO's Regional Office for Africa and with colleagues from Africa and elsewhere to develop an HSR training and small-grants program and to publish the *HSR Training Course* that was a forerunner of this volume. She currently works as a consultant in international health for groups such as WHO, IDRC, and Wellstart and teaches at the University of California at San Diego.